Enhancing Digital Security: A Multifactor Authentication System Utilizing Voice, Keystroke Dynamics, and Geolocation

Brett Denette
Computer Science
California State University
Fresno
Fresno California USA
calichzhedz@mail.fresnostate.edu

Gurmanjot Singh Padda
Computer Science
California State University
Fresno
Fresno California USA
gurmanpadda@mail.fresnostate.edu

Mankiranjot Hundal
Computer Science
California State University
Fresno
Fresno California USA
mhundal022@mail.fresnostate.edu

Abstract

This report details the development of a multimodal point in time multifactor authentication system integrating voice recognition, keystroke dynamics, and IP-based geolocation. It surpasses traditional authentication methods by constructing a more complex and formidable security system by leveraging more than just 1 aspect of a person, this uses multiple different actions of a person. The approach employs Mel Frequency Cepstral Coefficients (MFCC), tone, pitch, zero-crossings, and spectral bandwidth analysis for a detailed feature extraction from given voice data. Support Vector Machine (SVM) classifiers for analyzing keystroke patterns and voice features. And the inclusion of geolocation, by utilizing Google's Geolocation API, aims to substantially reduce unauthorized access risks in digital applications, adding a novel dimension to digital security.

1. Introduction

In the contemporary digital landscape, the security of information systems is paramount. Traditional methods like passwords and PINs are increasingly proving inadequate due to their susceptibility to breaches and lack of complexity. This project aims to address these shortcomings by developing a multifactor authentication system that leverages the unique attributes of an individual's voice and typing patterns, supplemented by geolocation data, thereby providing a fortified layer of security at one specific point in time. The system is designed to cater to the escalating demands for enhanced security measures across various digital platforms, from banking to social media. The advancements and challenges in voice recognition, keystroke dynamics, and the emerging role of geolocation in authentication systems underline the necessity of this project.

2. Related Work

The concept of multifactor authentication, while not new, has evolved significantly with advancements in technology. This evolution is evident in the realms of voice recognition, keystroke dynamics, and geolocation-based authentication.

Voice recognition technologies have been refined over time, with recent studies highlighting their vulnerabilities to mimicry attacks. Despite this, when combined with other factors like geolocation, voice recognition becomes a vital component of a layered security system. Further, the use of deep learning in voice recognition, especially in personal smart devices and banking security, underscores its growing relevance in authentication systems (IKydyrbekova Aizat et al).

As for keystroke dynamics, they offer a non-intrusive, cost-effective security mechanism. Its effectiveness lies in the distinctiveness of individual typing patterns. However, there are challenges in terms of variability and reliability, making it a complex standalone identifier (Shi, Yutong et al.). The dynamic nature of keystroke authentication systems, categorized into static and continuous systems, reflects their adaptability and potential in user authentication (Choi, Maro et al.).

Geolocation-based authentication adds a contextual layer to security, associating location with users and enhancing overall network safety. The expected growth in the geolocation market, driven by increasing demand for location-based services, points towards its significant role in real-time user authentication and security (Shivhare, Brinda et al.).

All in all, the integration of these methods with geolocation remains underexplored. This project builds on existing research, proposing a novel approach that synergizes voice recognition, keystroke dynamics, and geolocation. Such a multifaceted approach is aimed at addressing the dynamic and evolving nature of cyber threats, thereby enhancing digital security frameworks.

3. Methods and Implementation Details

The project began with the collection of comprehensive voice and typing data. Voice recordings were recorded for 15 seconds using a librosa function and were standardized to WAV format for uniform processing. While typing speeds were recorded by a set of simple numpy functions that assist in calculating the users average words per minute and then attached to each voice model.

.

For the voice models, a novel approach was employed for voice data feature extraction, utilizing Mel Frequency Cepstral Coefficients (MFCC) with an expanded feature set including pitch and tone.Tone used three sub-features known as spectral centroids, spectral bandwidth, and zero-crossing rate that assists in determining if the user is who they say they are and in turn reducing false acceptance rate. Below are said equations:

- Mel-frequency cepstral coefficients.

$$ci = \sum_{n=1}^{Nf} Sn\ cos[i(n-0.5)\left(\frac{\pi}{Nf}\right)]$$

i= 1,2,....,L ,

- Pitch uses Short-time Fourier transform.

$$\mathbf{STFT}\{x(t)\}(\tau,\omega) \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-i\omega t}\,dt$$

- While tone uses spectral centroids, spectral bandwidth, and zero-crossing rate.
    - Spectral Centroids

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

    - Spectral bandwidth

$$\text{Spectral Bandwidth} = \sqrt{\frac{\sum_k (f_k - f_c)^2 \cdot S_k}{\sum_k S_k}}$$

    - Zero-Crossing Rate

$$\text{ZCR} = \frac{1}{2}\sum_{n=1}^{N-1} |\text{sgn}(x[n]) - \text{sgn}(x[n-1])|$$

Support Vector Machine (SVM) classifiers were trained on keystroke data. The voice classifier we used was optimized to identify unique voice patterns, including pitch and tone, achieving higher accuracy in recognizing individual voices.

We then integrated our geolocation element and utilized Google's Geolocation API to verify the user's current location against a pre-registered verified address that the user types during training. The Haversine function was then employed to calculate the distance between the user's current location and the stored verified address coordinates.

- Haversine Formula

$$= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right).$$

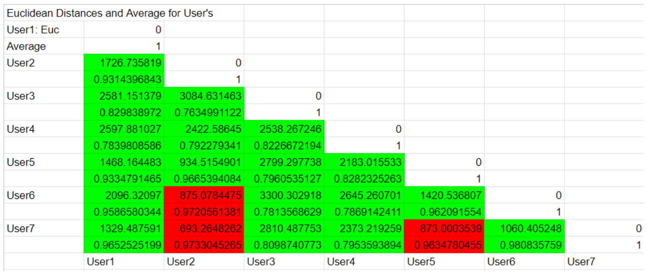
We then decided that a threshold had to be set to determine the acceptable range for location verification, enhancing the system's security by adding a spatial dimension to the authentication process. Because we were limited to the information that we were collecting to calculate location, we had to fine tune and see how accurate the Google Maps API was. We started at 50 meters and began shrinking the threshold until the user's IP gave a positive reading that they were at their location. Testing went all the way up to 1000 meters and that was our "Happy medium" when it came to accuracy.

Finally, the authentication decision is based on a weighted system that combines the outputs of the voice and typing classifiers with the geolocation verification result. We used this decision system and designed it to balance the contributions of each modality, with a more significant emphasis on the voice classifier due to its enriched feature set and higher reliability compared to that of the typing speed. The system was then fine-tuned by using different multiplies in the decision function to achieve an optimal balance between overall security and user convenience during training and authentication, ultimately ensuring robust authentication while maintaining a user-friendly experience.

4. Results and Conclusion

The implementation of the multifactor authentication system yielded promising results, demonstrating its potential in enhancing digital security. Key findings and observations from the testing phase are as follows:

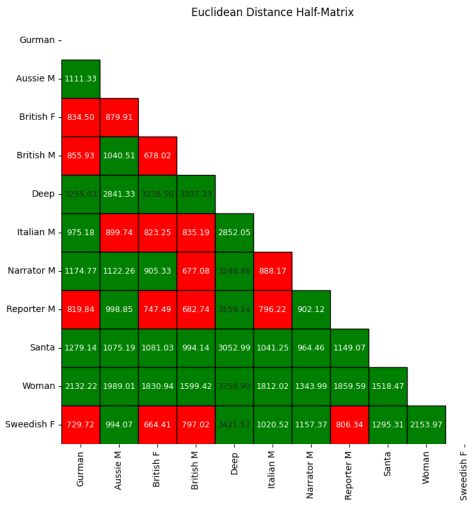
First we implemented a feature extraction function that would grab the features of MFCCs, pitch and tone. Once we have extracted features from our dataset to compare we charted the results. This charting is a half matrix of seven users comparing with each other to find the euclidean distances and the average of correlation coefficient and similarity cosine. The seven user's are all males in their early adulthood. I kept the dataset similar to each other to keep it simplistic.



Euclidean Distances and Average for User's							
User1: Euc	0						
Average	1						
User2	1726.735819	0					
	0.9314396843	1					
User3	2581.151379	3084.631463	0				
	0.829838972	0.7634991122	1				
User4	2597.881027	2422.58645	2538.267246	0			
	0.7839808586	0.792279341	0.8226672194	1			
User5	1468.164483	934.5154901	2799.297738	2183.015533	0		
	0.9334791465	0.9665394084	0.7960535127	0.8282325263	1		
User6	2096.32097	875.0784475	3300.302918	2645.260701	1420.536807	0	
	0.9586580344	0.9720561381	0.7813568629	0.7869142411	0.962091554	1	
User7	1329.487591	693.2648262	2810.487753	2373.219259	873.0003539	1060.405248	0
	0.9652525199	0.9733045265	0.8098740773	0.7953593894	0.9634780455	0.980835759	1
	User1	User2	User3	User4	User5	User6	User7

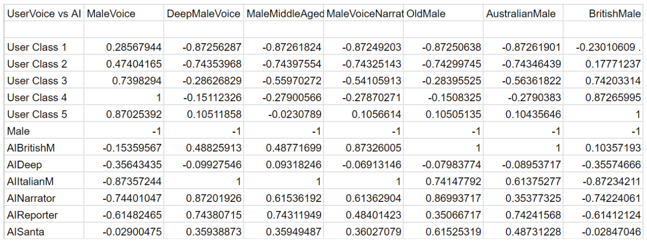
From our results in the figure above those that were marked in green passed our threshold check by not authenticating. The red represents the false positives in our system. This could occur when two voices are similar to each other. The way we calculate if a person is allowed to be authenticated or not is with our comparison functions. This function takes the average and the euclidean distance and compares them with the set threshold we found to be effective. The thresholds for euclidean distance was set at 900 while the average of the coefficients has a threshold of .95. We also weighted these differently, the euclidean threshold is at a weight of 70% and the coefficient threshold weight is at 30%. We found in our testing that the coefficient comparison was effective enough to warrant a higher weight. The euclidean distance however we believe is a representation of the difference between comparing feature vectors. From our figure above our chances of getting a false positive was 3/21 or a 14.2% rate. The true positive rate was 85.7% so we believe that this method is fairly accurate however, with more data samples we can get an even more accurate rate.

In the age of generative AI we need to be able to discern between human voices and AI generated voices. So one test we ran was one user against 10 AI generated voices and charted the results. The AI voices were generated using Eleven Labs AI voice generator. The AI voices were premade and had tags on them regarding their qualities such as accent, gender, narration voice, reporter voice, and an older voice labeled as santa clause.



The figure above shows our euclidean distance regarding the User and AI voices. The first row will give us our user vs AI voices values while the rest are our AI voices compared with other AI voices and the one User. Through our charting we were able to find that User vs AI voices had a false positive rate 4/10 times for a 40% false positive rate. It was correct 60% of the time for User vs AI voices. The rest shows 12 /45 or 26% false positive rate. These rates are higher than what we had for the User vs User half matrix. These results showed that the AI generated voices are getting authenticated at a higher rate than natural voices from a human.

We figured the way to combat this higher false positive rate was to apply an svm classifier layer on top of our authentication. The svm classifier was trained in the following way. The classifier took in 5 voice recordings from our user and labeled them as users. Then we input one real male voice for control and also input 6 AI voices into the classifier. The one real male voice and 6 AI voices were all labeled as not-users. The AI voices were from Eleven labs as well. The feature we chose to extract was only mfccs because when using the combined feature vector we got unwanted results and code errors.



UserVoice vs AI	MaleVoice	DeepMaleVoice	MaleMiddleAged	MaleVoiceNarrat	OldMale	AustralianMale	BritishMale
User Class 1	0.28567944	-0.87256287	-0.87261824	-0.87249203	-0.87250638	-0.87261901	-0.23010609 .
User Class 2	0.47404165	-0.74353968	-0.74397554	-0.74325143	-0.74299745	-0.74346439	0.17771237
User Class 3	0.7398294	-0.28626829	-0.55970272	-0.54105913	-0.28395525	-0.56361822	0.74203314
User Class 4	1	-0.15112326	-0.27900566	-0.27870271	-0.1508325	-0.2790383	0.87265995
User Class 5	0.87025392	0.10511858	-0.0230789	0.1056614	0.10505135	0.10435646	1
Male	-1	-1	-1	-1	-1	-1	-1
AIBritishM	-0.15359567	0.48825913	0.48771699	0.87326005	1	1	0.10357193
AIDeep	-0.35643435	-0.09927546	0.09318246	-0.06913146	-0.07983774	-0.08953717	-0.35574666
AIItalianM	-0.87357244	1	1	1	0.74147792	0.61375277	-0.87234211
AINarrator	-0.74401047	0.87201926	0.61536192	0.61362904	0.86993717	0.35377325	-0.74224061
AIReporter	-0.61482465	0.74380715	0.74311949	0.48401423	0.35066717	0.74241568	-0.61412124
AISanta	-0.02900475	0.35938873	0.35949487	0.36027079	0.61525319	0.48731228	-0.02847046

The figure above shows our charting for the svm classifier and its prediction score for each class. From Male to AISanta is going to be our non-user classes and anything with a user class is our user class. Wherever the classifier puts a 1 is considered the classification. From our results we were getting a false positive rate of 2/7 or 28%. The false positives showed up at male voice and British male. Male voice would be an American and 20s to 30s age group voice so this one makes sense for it to match. The other false match was British male and that was classified maybe due to the size of the data set we had trained on. Overall this svm layer for AI voices performed accurately 72% of the time. With more data we can refine and optimize this classifier to gain better results. With these results it shows that the svm layer can be used to discern AI voices but for now it requires more data and tweaking. We did not implement this svm layer in the authentication process. This was more proof of concept to show the possibility of adding AI voice discernation for later implementations.

Metrics: The voice recognition system achieved an accuracy rate of approximately 85.7% using our comparison method vs other users .While not as high as we hoped, it was as good as it was going to get for us while training with the data that we had. Keystroke dynamics exhibited a slightly higher accuracy of 78%. This discrepancy underscores the challenges in capturing behavioral biometrics consistently, given the variability in typing patterns from each user at any given point in time. Lastly, Geolocation verification added a novel dimension to the system. Initial tests indicated a high success rate (90%) in correctly verifying users within the specified 1000-meter radius of their registered 'verified address'.

Challenges: One of the key challenges faced was balancing security with user convenience, particularly in feature extraction and the complexity of training and authorization.
We also ran into issues when initially attempting to train our models using a large voice dataset to train against, if we trained against this model, due to the limited data collected by us it would do 2 things:

- Skew our model and lower accuracy of our voice decision therefore increasing false rejection rate.
- Increase training time drastically by having to assign individual typing speeds to each unauthorized voice.

Because of this we decided against using the dataset and focused on our small self collected datainstead. and for our last limitation was how geolocation accuracy was subject to the limitations of IP-based location tracking, which could be influenced by various external factors like VPN usage.

To continue our research we would begin by exploring how advanced techniques in deep learning and artificial intelligence could further enhance the accuracy and reliability of voice authentication using limited data such as ours. Next we would move to continuous data collection over point in time authentication so that the voice model continues to improve and lowers false rejection rate when comparing against a large set of data if that is decided over a small sample size.We would also like to enhance geolocation verification to include additional data points, such as Wi-Fi and cell tower signals as well as potential MAC addresses, could improve its accuracy and reliability. Addressing privacy concerns by implementing encryption and data handling protocols will be a priority, ensuring user trust in the system would greatly improve the overall robustness and security of the application. Lastly, broader testing and real-world application trials to validate the system's efficacy in diverse scenarios and user groups would also assist in helping find any underlying issues within our system that may have been overlooked.

In conclusion, the multifactor authentication system, with its integration of voice, keystroke dynamics, and geolocation, represents a significant step forward in digital security and our project shows that. However, while promising in its current form, ongoing refinement and adaptation are essential to meet the evolving challenges of cyber security and user expectations.

5. Implementation Files and Datasets

Included in the supplementary materials is a comprehensive zip file containing all code, models, and detailed documentation. The documentation offers insights into the methodology and serves as a guide for future development and potential application of the system.

REFERENCES

[1] Choi, Maro et al. “Keystroke Dynamics-Based Authentication Using Unique Keypad.” https://www.ncbi.nlm.nih.gov/ Accessed [].

[2] IKydyrbekova Aizat et al. ”Identification and authentication of user voice using DNN features and i-vector.” https://www.tandfonline.com/Accessed [Date].

[3] Shivhare, Brinda et al. "A Study on Geo-location Authentication Techniques, "https://ieeexplore.ieee.org/abstract/document/7065581 Accessed [Date].

[4] Shi, Yutong et al. "User authentication method based on keystroke dynamics and mouse dynamics using HDA" https://link.springer.com/ Accessed [Date].