A Combative Dispute on Security Framework Model of Mel Scale Mixed Methods in Speech Recognition System

Heni Ispur Pratiwi[1,*], Iman Herwidiana Kartowisastro[2], Benfano Soewito[3], Widodo Budiharto[4]

[1]*Doctorate Program of Computer Science, Bina Nusantara University, Jl. Raya Kebon Jeruk No.27, West Jakarta 1530, Indonesia*

[2,3]*BINUS Graduate Program, Master of Computer Science, Doctorate Program of Computer Science, and Computer Engineering Dept. Faculty of Engineering, Bina Nusantara University, Jl. Raya Kebon Jeruk No.27, West Jakarta 1530, Indonesia*

[4]*Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. H. Syahdan No.9, West Jakarta 11480, Indonesia*

(Received: February 14, 2025; Revised: April 11, 2025; Accepted: May 9, 2025; Available online: June 2, 2025)

Abstract

An audio recording system device has unprecedented activities of its authorized users which in a particular way cause vulnerability to the system. It starts to get into a fuzzy condition and deteriorate the system sensitivity in detecting unauthorized access to pass through, then the system inclination may occur. One case is when separate users picked speech voices with similar keywords to set their usernames or password. Moreover, when users are siblings or twins that could have merely similar voices. Troublesome of this situation leads to a less sensitive manner of a security system, and in some situations, the system could operate blocking authorized users themselves to get access. This paper defines a proposed method to resolve the situation by combining Mel Frequency Cepstral Coefficient with other methodologies, which have been implemented for many other research' specific objectives as well. This paper displays to prove its combination with an interval scoring in Fuzzy Relation complements a resolution to tackle the security of fuzzy issues mentioned. The Mel Scale has its capacity of delivering extractions output from audio input data, it is called as spectral centroids which refer to humans' voices or an individual's voice features. Some spectral centroids get merely similar results due to those inclinations mentioned. This paper exposes Fuzzy Relation method to fit the need of verification procedures thorough its interval scale on any fuzzy features. The objective of verification procedure is to gain consistency measured scales, and security warrant remains valid. The inhouse experiments served to give user A of [0.49, 1.18] interval, user B of [0.76,1.07] interval, and user C of [0.44,0.95] interval, and those interval numbers are proposed to cap other login users accounts unto theirs.

Keywords: Spectral Centroid, Mel Scale, Fourier Transform, Fuzzy, Speech Character

1. Introduction

Extending the discussions of speech recognition system (SRS), speech voices feature consistency are captured by utilizing Mel-Scale dan Fourier Transform [1]. The article had conducted an experiment with a proposed methodology model of mixed Mel Frequency Cepstral Coefficient (MFCC) and Short Time Fourier Transform (STFT). The proposed customized username and passwords based on unique character of user's speech voices in significant length of time distances. Consistency of spectral centroids should be capable to block unauthorized access. A conditional dispute could pop up on spectral centroids consistency with unprecedented conditions. One challenge condition to the system is when the unauthorized access is unintended and its undermining tracks are not visible. The spectral centroids consistency procedures could be overlooked when merely similar voices from similar gender of siblings or identical twins' role as users at one single electronic device.

Moreover, they applied similar key words to customize their usernames or passwords causing a fuzzy condition figure 1. Users' contentious security manners are the facts of vulnerability exposures to the system which open loopholes to any illegal accesses, such as recorded speech voices activities from authorized users which eventually embrace a fuzzy condition and it gradually irritates sensitivity of the security system. The objective of this paper is eagerly to deliver a dispute to the existed framework, as mentioned, by exposing this unprecedented fuzzy condition. Refer to the human

***Corresponding author: Heni Ispur Pratiwi (heni.pratiwi001@binus.ac.id)**

DOI: https://doi.org/10.47738/jads.v6i3.689

anatomy system have organic nature elements, that carry some rewards by defaults, to an individual unique character and it has measurable sensitivities [2]. The dispute brought further observations on speech voices character consistency that should remain significant and critical scope topic along through this paper.

In the objective context, the need of combining MFCC with another methodology is once more required to offer a resolution on the dispute mentioned. The MFCC extracted features from the selected dataset should prove that the appointed dispute at the existed system remains solvable. Regarding the selected data should display identical gender speech voices with similar spoken words, this is to meet the criteria of the dispute mentioned. The identical gender criteria are considered on getting close of biological anatomy realms [2].

2. Literature Background

Generally, a speech recognition system (SRS) has an operational mechanism set to have speech commands from authorized users, it serves for activities such as recordings, editing, and savings. One common activity is to set and reset login keys which are verified to claim speech voice commands into the device control system. In the situation of a single device has overwhelmed with variety users' login keys could compromise the system, in some cases, to become less sensitive detecting differences and false identifying illegal access of unauthorized users and cause the control system access compromised. One of the case examples is the dominant spectral speech and prerequisite environment combined to affect several security cases, especially in siblings or identical twins most likely to have resemblances in producing their voice tones, accents, and frequencies [11], [12], [13], [14]. Moreover, those identical resemblances are affected from a vast variability of signals and expose an ambiguity possibility due to the facts that the same speaker may speak one particular word differently. The fact of different spelling or pronouncing on particular words are called as speaking style shifts [13], [15], [14], [16], [18]. Besides, many new technologies had updated recording devices apps and its transmission process on top of the existed complexity of the issues.

The idea of origin in exposing STFT capacity into finding speech voices character within time series interval duration, the components of each character are displayed with dynamic scale of changes due to human ages [19], [20]. The experiment showed that those characters are represented by spectral data found within one, or more, time intervals. In facts, Fourier conventional analysis will be supportively plausible when there was no spectral to time domain conversion process [21], [22]. Since speech voices tie to continuous time-varying signals with challenging rate changes every 10 to 30 times/seconds [22], at this point shows that MFCC needs other methods to complete the objective. In the case of previously mentioned on how STFT works in finding components or elements to identify speech voices characters with specific manners, and brings a visible comprehensive system to profile those characters to be built with more modified and expanded procedures for security concerns.

There are four components to base an SRS. First, the system must have a sensor module. Second, the system must have feature extraction module. Third, the system must have matching modules, and fourth, the system must have decision-making modules to accept or reject user's login input [21], [23]. Commonly, those four components are dominant parts of voice recognition techniques for Automatic Speaker Verification and Automatic Speaker Identification. Initially users are located at enrolment time which are required to create login keys with short phrases and repeat them to confirm. The confirmation is expected to shield the illegal access from unauthorized users.

3. Methodology

Security authentications are widely enhanced through implying variety data sources which means it claims for deeper observations of utilizing other methodologies to fit on the objective needs. In general, selecting methodologies due to data environment and research objective output and in the context of the dispute mentioned. Hence, secondary dataset was selected from URL: https://www.kaggle.com/kongaevans/speaker-recognition-dataset. The dataset consists of three male audio speech voices who have similar keywords and they are repeatedly spoken. It is expected to take its chance to utilize them in unfolding the ambiguity manner in the system as what the dispute mentioned. Topping up the existed combined methods (table 1), here an application of FR onto MFCC features is to obtain scores in resolving fuzziness and identifying particular users.

Combined MFCC with other methodologies had shown significant successes, table 1 shows that it had been implemented by many of researchers. The significant starting point was stated by Gyulyustan on MFCC and Linear Predictive Coding (LPC) [3], continued by other researchers displayed on table 1, researchers gave the perspective of adopting other methodologies to support the need of Speech Recognition Security (SRS) facts. Bibliography references welcome to explore about more perspectives shown at table 1. Synchronized facts and its elements are spotlighted spaces for researchers to explore more experiments with more combined methodologies.

Table 1. Combined MFCC and other methodologies

No.	Methodologies	References
1.	MFCC, SVM, MLP	[4]
2.	DTCWT, MFCC	[5]
3.	MFCC, ENTROCY	[6]
4.	MFCC, DTW	[7]
5.	MFCC, Random Forest	[8]
6.	MFCC, Baum Welch	[9]
7.	MFCC, LPC, CNN	[10]

Some particular methodologies could be used to gain spectral features in advance of implying fuzzy relation for scores, however, MFCC methodology is used for the context of issues in this paper [24]. The basic reason of selecting MFCC base on its capacity of delivering extractions output from audio input data which in this particular case is the spectral centroid. The spectral centroid refers to the facts of humans' voices or an individual's voice feature based on physical characteristics in creating a sound, such as mouth and lips, nasal cavities and vocal tracts [13]. The following figure 2, which is the extension of figure 1 and it is to illustrate in general when some users have similar usernames or passwords for login keys in a single SRS device.

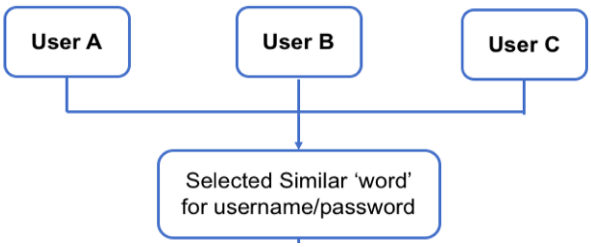

Figure 1. A common user activity could lead to fuzz the system.

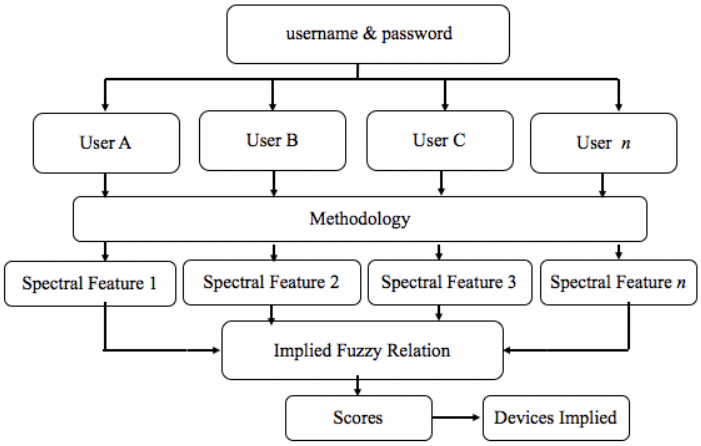

Figure 2. The proposed mechanism designs.

The Selection of dataset are also considered to observe inclination conditions, and the experiment remain to have an objective in obtaining the consistency value of individual speech voices character. The similar spoken word of "zero"

by four different males repeatedly are captured to measure individual spectral centroid. The experiment of combining MFCC with FR is objected to resolve a dispute of ambiguity due to inclination causes. Figure 3 shows how those repeatedly spoken similar words are being rolled to get validated with the terms of (1) to (5). The background of figure 3 exposure on the inconsistency results from multiple users with repeatedly spoken similar words lead to conducted series of experiments on recording three male speakers that were randomly required to say the same ordered words of "some 12 or more" repeatedly [13]. The experiment had discovered that every speaker owned varied actual frequency measurements along the tests. The results had observed about the frequencies of the first and second formants, they showed consistency ratio. However, when those two formants were plotted in a two-dimensional plane against one another, repetitions of each vowel sound seemed like to create a distinct cluster. Furthermore, Potter and Steinberg delivered the facts that individual uniqueness remain significant despite similarity [13]. This paper is inspired to have slightly extended from what Potter and Steinberg findings, which the distinct clusters output is adopted to score individual uniqueness for a security complement.

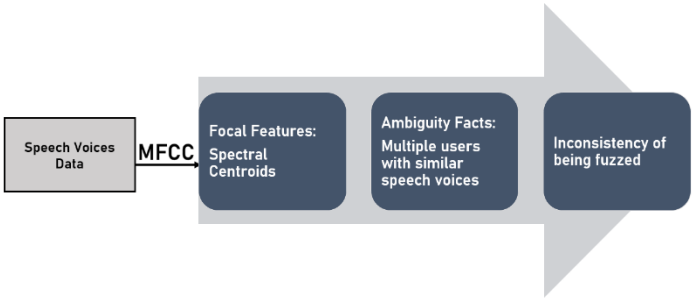

Figure 3. Methodology of MFCC workflow outline.

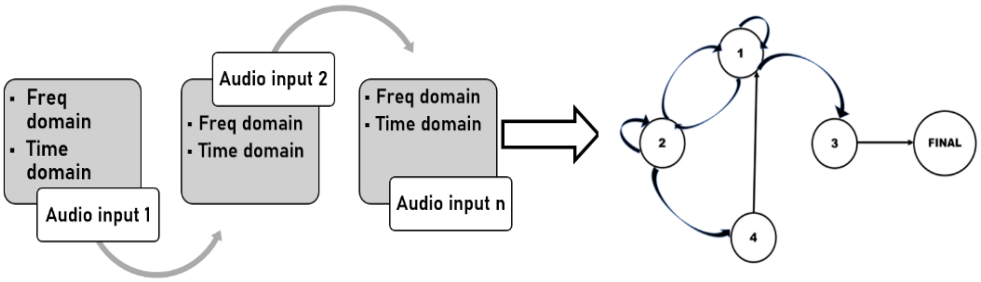

Figure 4. FR mapping workflow outline.

Figure 4 displays how repeatedly spoken words are captured as audio input 1 and so on. Any audio input has its measured frequency and time domain, which are the MFCC feature output. Furthermore, the derivatives values are simply rolling its validations through the final. This visualization would give supportive ideas in utilizing MFCC features.

3.1. Mel Frequency Cepstral Coefficient (MFCC)

MFCC remains as a commonly used method for feature extractions, especially dealing with a design of SRS [19], [21], [24], [25], the extraction technique simply applies using 1024 frequency bins, a frameshift and fixed time length in between 20 to 32 Ms, 26 Mel channels, and 10 to 40 cepstral coefficients (with normalized cepstral mean).

The frameshift and time length are destined to gain stable parameters, the time length of 20 to 32 milliseconds are considered to be sufficient to examine speech signal data. This phase in MFCC is called framing or windowing stage. Those speech signal data input are not granted to be clear from distortions, and Hamming window phase serve to reduce the distortions. Since MFCC has the 1024 frequency bins, there is a need to do time domain windows conversion into frequency domains to form them into a regular grid by discretizing and interpolating the windows. Hence, Fourier transform is smoothly applied. The Mel filter provide channels to make the selective human auditory system works and returns coefficients as the output [19], [20], [24].

The SRS must be trained to the individual's voice at enrollment time, and more than one enrollment session is often necessary. Then feature extraction module mostly measure unique formants or sound characteristics from the individuals' vocal tract. Those formants have its frequency positions and create significant patterns to discriminate the speech voices [20], [21], for example, when specific vowel sounds are spotlighted as distinct facts from one speaker to another. The matching module compares those facts using the pattern algorithm, and the decision-making module states the acceptance or rejection due to its security system output conventionally. In addition, assigned interval scores to each individual significant patterns of speech voices unto the system which is expected to top up the existed security system and initiate the access permission into the system to control. This paper displays a practical method to obtain the scores by applying Fuzzy Relation.

In advance of spectral centroid descriptions, there is a need to state about a speech voice. A human individual's voice feature is used and based on physical characteristics in creating a sound, such as mouth and lips, nasal cavities and vocal tracts. These features shape human speech characters that are invariant for an individual, but over time due to age, health conditions and emotional state cause behavioral changes. However, behavioral traits could serve as individuals' uniqueness when they are common or universal for people to have it, and it should be distinctive and permanence [13], [19].

3.2. Fuzzy Relation

There are some distinct protocols between Fuzzy Theory, Fuzzy Logic, and Fuzzy Relation [26]. Bibliography lists attached are pleased to lead their details. This paper delivers a brief description on Fuzzy Relation (FR) in its context of utilizing it as an additional metric on security matters. The description of FR ticks to come with the term of 'relation' literally defines mathematical concept of mapping related variables which are in one set or group. The term of relation is used to measure each member or element dependencies to the others in the group or dependencies between one group to the other group. In fact, a dependency model is patterned from around this context.

FR is a system contained of a fuzzy set with the rules of the Cartesian product crisp definition which has variant memberships within its structure, and they are related to one another due to some respect from other member forms with similar values or manners from a set of standards [27]. In the case of Crisp Relation, there are two tiers degrees of mapping the dependency between elements, either they are completely-related or not-related. The result is absolute as correct output. In FR, however, the mapping degrees have infinite-number values. The mapping degrees have developed with more varieties, and some degrees between two or more elements are called as extreme positions of being completely-related or not-related [13]. This extreme position is to advantage a procedural security verification.

FR adds to adopt a supportive resolution when data varieties cause the system output shows compromised. The variant memberships are in real numbers form of interval [0,1] which indicates the close or strong relation between the members within the interval structure itself [28]. Equation (1) shows the practical implications of the FR on SRS, $X(n)$ represents users, and (2), $x(n)$ represents each user's speech voice extractions. As for (6) and (7), the interval [0,1] is read as [minimum, maximum] and obtained from overall data.

$$X(n) = X1, X2, X3, \dots, Xn \quad (1)$$

$$X(n) = X1, X2, X3, \dots, Xn \quad (2)$$

$$y = f(x(n)) \quad (3)$$

$$y' = f'(x(n)) \quad (4)$$

$$y'' = f''(x(n)) \quad (5)$$

$$y'' > 0 \; then \; y = minimum \quad (6)$$

$$y'' < 0 \; then \; y = maximum \quad (7)$$

4. Results and Discussion

In SRS, security manners capture perception of human speech voices using the dominant frequencies or spectral peak positions in speech signals [18]. There are some facts and challenges in measuring a security system of SRS

performance, beside the socio linguistic, gender and age which are dominant facts of spectral speech signal peak positions [18], [29]. SRS technology, in some applied rules, the voice commands are set to use limited words due to prerequisite environment [29]. In facts, there are some standards rules to characterize an individual voice which have been developed from SRS technology field. One of the rules states that signal bandwidth has a standard of 4 kHz, while signal frequencies in periodically travelling waves like from 80 Hz to 350 Hz. The periodical waves have time series executions in certain points inclusively, which are called as spectral energy distributions, especially when a peak frequency occurs [20], [21].

The term of spectral energy distributions refers to the vowel channels, which could change to be better after articulations trainings. The term of spectrograms represents three dimensions of speech voice spectral signals, in horizontal axis as time domain and vertical axis as frequencies, which vary from one speech to the others. The up and down of human voices spectrum normally occur in every octave of 6 dB, which relates to the individual speech voices character [20]. Considering those related points, this paper serves to display observation through published journals and conference articles, and implying experiments to resume that some ambiguities are still resolvable. Following figures are in house experiments to support the proposed resolution. Figure 5a below shows how user A scores numeric are gained from the extractions results of MFCC and its maximum of 1.18 and minimum of 0.49 computations displayed as bar charts in figure 5b below:

1,22	1,12	1,15	1,10	1,05	1,17	1,01	0,98	1,02	0,93	1,06	1,05	0,87	1,18	1,06
1,01	0,97	1,09	0,92	0,97	1,06	1,07	1,02	1,01	0,93	0,94	0,96	1,07	0,96	0,92
1,11	1,14	1,09	1,10	1,09	1,04	1,12	1,07	1,08	1,12	1,21	1,02	0,96	0,96	

Figure 5a. User A numeric data

Bar charts of figure 5b are to accommodate the computed extractions data of 5a and to display user A scores of each input. Here is figure 5b below:

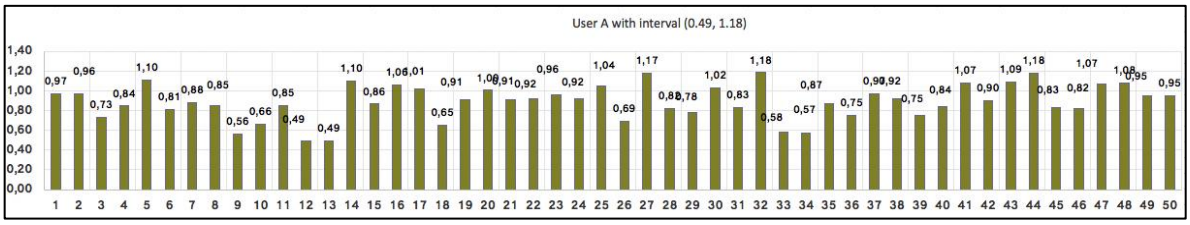

Figure 5b. User A bar charts data with interval (minimum: 0.49, maximum: 1.18)

Figure 6a below shows how user B scores numeric are gained from the extractions results of MFCC and its maximum of 1.07 and minimum of 0.76 computations displayed as bar charts in figure 6b below:

0,97	0,97	0,89	1	0,83	0,92	0,82	1,07	0,92	0,8	0,97	0,93	0,88	0,88	0,89	0,93
0,89	0,9	0,92	0,85	0,94	1,03	0,97	0,81	0,86	0,82	1	1,01	0,89	1,05	0,77	0,84
0,89	0,84	0,89	1,05	0,97	0,94	0,92	0,8	0,93	0,94	0,82	0,76	0,91	0,86	0,93	0,95

Figure 6a. User B numeric data

Bar charts of figure 6b are to accommodate the computed extractions data of 6a and to display user B scores of each input. Here is figure 6b below:

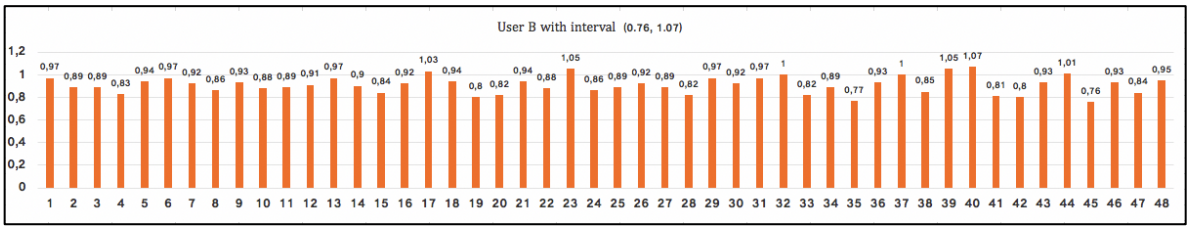

Figure 6b. User B bar charts data with interval (minimum: 0.76, maximum: 1.07)

Figure 7a below shows how user C scores numeric are gained from the extractions results of MFCC and its maximum of 0.95 and minimum of 0.44 computations.

ISSN 2723-6471

0,80	0,86	0,82	0,44	0,71	0,80	0,71	0,69	0,60	0,60	0,69	0,80	0,86	0,72	0,79	0,53	0,63
0,77	0,95	0,66	0,74	0,88	0,63	0,58	0,57	0,79	0,60	0,73	0,59	0,75	0,85	0,80	0,85	0,74
0,82	0,87	0,90	0,58	0,92	0,92	0,71	0,83	0,59	0,78	0,67	0,83	0,56	0,78	0,86	0,64	

Figure 7a. User C numeric data

Bar charts of figure 7b are to accommodate the computed extractions data of 7a and to display user C scores of each input. Here is figure 7b below:

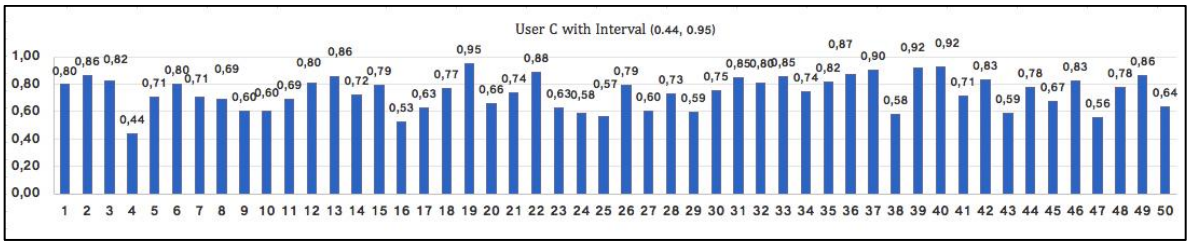

Figure 7b. User C bar charts data with interval (minimum: 0.44, maximum: 0.95)

Figure 5a, 6a, and 7a show how scores numeric are gained from the extractions results of MFCC and its maximum and minimum computations. Bar charts are displayed as figures 5b, 6b and 7b are to accommodate the computed extractions data of figure 5a, 6a and 7a and to display individual's score of each user. From those bar charts displayed, the results found that the interval numeric at minimum value of user A (figure 5a and 5b) and C (figure 7a and 7b) are close by 0.05 point at the maximum values, and each user has showed to be nearby 0.1 to 0.2 scores. Here, the scores show spectral peaks positions and spectral energy distributions for each user, and those are captured as their speech voices characters. Though signal of speech voice is periodically, travelling waves like, from 80 Hz to 350 Hz, their speech voice characters should be consistent due to its uniqueness.

5. Conclusion

Refer back to the first paragraph, the MFCC features are to complement STFT in finding elements or components to characterize individual's speech voices in specific interval of time series. The findings are proposed to be applied as username/passwords selections for the particular users. This paper comes to extend a dispute situation when those proposed elements found are applied or selected by users who are most likely to have resemblances in producing their voice tones, accents, and frequencies, such as among siblings or identical twins. In the other hand, a spatial ambiguity occurs when there is no sufficient provision to represent each region of uncertainties [27]. In fact, the relations of its features are necessarily needed to retain information for highest level decision.

In such a fashion, the experiment result with selected dataset would not be appealing to meet and fit the dispute mentioned. Selected dataset will not represent what exactly is required to have identical twin or close sibling relations. However, for such some other perspective fashion, the experiment result is once again surfacing resolvable issues of ambiguity manners in the system environment. Those interval values from the charts could be appealing and satisfying when the required dataset to fit exactly on the dispute context.

Feature extractions are essentially important in speech signal for speech recognition. The MFCC remains to be a popular method to accomplish the task. Those features are presumed to get utilized for a security system objective. In this case, the manifestation of spectral centroid features are applied for a security assessment. The practical assessment is simply to track users' activities that create an ambiguity or fuzzy condition and cause the system to be less sensitive, and passing the control access for unauthorized users. Consistency of speech voice character remain to be obtained, despite of the resemblances of voice tones, accents, and frequencies. A larger dataset could be more baneful to gain accuracies and efficiency measures, and it will give more detected output patterns and insight into system adaptability. So, any fault on implied methodology could be re-designed and developed. Implications of FR as a security perspective in this paper is just a beginning, there are more FR properties that could expand future researches to enrich security topics.

6. Declarations

6.1. Author Contributions

Conceptualization: H.I.P., I.H.K., B.S., and W.B.; Methodology: H.I.P., and I.H.K.; Software: H.I.P.; Validation: H.I.P., I.H.K., B.S., and W.B.; Formal Analysis: H.I.P., I.H.K., B.S., and W.B.; Investigation: H.I.P.; Resources: I.H.K.; Data Curation: I.H.K.; Writing—Original Draft Preparation: H.I.P., I.H.K., B.S., and W.B.; Writing—Review and Editing: I.H.K., H.I.P., B.S., and W.B.; Visualization: H.I.P. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H.I. Pratiwi, I.H. Kartowisastro, B. Soewito, and W. Budiharto, "Short Time Fourier Transform in Reinvigorating Distinctive Facts of Individual Spectral Centroid of Mel Frequency Numeric for Security Authentication," *International Journal of Innovative Computing, Information and Control (IJICIC)*, vol. 20, no. 02, pp. 213-227, 2024.

[2] D.R. Kisku, P. Gupta, and J.K. Sing, "Advances in Biometric for Secure Human Authentication and Recognition," *CRC Press, Taylor and Francis Group*, vol.2014, no. 01, pp. 3-12, 2014.

[3] H. Gyulyustan, H., and S. Enkov, "Experimental Speech Recognition System Based on Raspberry Pi 3," *IOSR Journal of Computer Engineering (IIOSR - JCE)*, vol. 19, no. 3, pp. 107-112, 2017.

[4] S.I. Khamlich, Atouf, F. Khamlich, and M. Benrabh, "Performance Evaluation and Implementations of MFCC, SVM and MLP Algorithms in the FPGA Board," *International Journal of Electrical and Computer Engineering Systems*, vol. 12 no. 3, pp. 139-153, 2021.

[5] H.C. Shantakumar, G.S. Nagaraja, and M. Basthikodi, "Performance Evolution of Face and Speech Recognition system using DTCWT and MFCC Features," *Turkish Journal of Computer and Mathematics Education*, vol. 12 no. 3, pp. 395 – 404, 2021.

[6] D.Y. Mohammed, K. Al-Karawi, and A. Aljuboori, "Robust speaker verifications by combining MFCC and Entrocy in noisy conditions," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2310-2319, 2021.

[7] B. Birch, C.A. Griffith, and A. Morgan, "Environmental Effects on Reliability and Accuracy of MFCC Based voice Recognition for Industrial Human-Robot-Interaction," *Proc IMechE Part B: J Engineering Manufacture*, vol. 235, no. 12, pp. 1939–1948, ImechE, 2021.

[8] V.S. Wicaksana, and A. Zahra, "Spoken Language Identification on Local Language Using MFCC, Random Forest, KNN, and GMM," *IJACSA Int. Journal of advanced of comp science and applications*, vol. 12 no. 5, pp. 394-398, 2021.

[9] M. Maseri and M. Mamat, "Performance Analysis of Implemented MFCC and HMM-based Speech Recognition System," *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia*, vol. 2020, no. 1, pp. 1-5, 2020, doi: 10.1109/IICAIET49801.2020.9257823.

[10] A. Chowdhury, and A. Ross, "Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals," *IEEE Transactions on Information Forensic and Security*, vol. 15, no. 1, pp.1556-6013, 2020.

[11] M. Malik, M.K.Malik, K.Mehmood, and I. Makhdoom, "Automatic Speech Recognition: A Survey," *Springer Science+Business Media*, LLC, part of Springer Nature, vol. 80, no. 3, pp. 1-47, 2020.

[12] T. Gunendradasan, B.Wickramasinghe, P. Ngoc Le, E. Ambikairajah, and J. Epps, "Detection of Replay-Spoofing Attacks Using Frequency Modulation Features," *Hyderabad: Interspeech 2-6*, vol. 2018, no.9, pp. 636-640, 2018.

[13] X. Li and M. Mills, "Vocal Features: From Voice Identification to Speech Recognition by Machine," *Technology and Culture*, vol.60, no. 2, pp. s129-s160, Johns Hopkins University Press, 2019.

[14] Y. Chen, X. Yuan, A. Wang, K. Chen, S. Zhang, and H. Huang, "Manipulating Users' Trust on Amazon Echo: Compromising Smart Home from Outside," *Endorsed Transactions on Security and Safety*, vol. 6, no. 22, pp. 1-12, 2020,

[15] J.S. Edu, J.M. Such, and G. Suarez-Tangli, "Smart Home Personal Assistants: A Security and Privacy Review," *ACM Computer Survey 1*, vol. 53, no. 6, article 116, pp. 1-12, August 2020.

[16] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: Unlocking Synthetic Data for Speech Recognition," *ISCA*, vol. 2021, no. 8, pp. 1-5, 2021.

[17] H. Voss, H. Wersing, and S. Kopp, "Addressing Data Scarcity in Multimodal User State Recognition by Combining Semi-Supervised and Supervised Learning," *ICMI*, vol. 2021, no.10, pp. 8–22, 2021.

[18] J.L.K.E. Fendji, D.C.M.Tala, B.O.Yenke, and M.Atemkeng, "Automatic Speech Recognition and Limited Vocabulary: A Survey," *Applied Artificial Intelligence*, vol.36, no.1, pp. 1-12, 2022

[19] H.I. Pratiwi, I.H. Kartowisastro, B. Soewito, and W. Budiharto, "Adopting Centroid and Bandwidth to Shape Security Line," *IEEE Xplore Conference Icosnikom*, Medan, Indonesia, doi:10.1109/ICOSNIKOM56551.2022.10034929, 2022.

[20] T. Safavi, "Automatic Speaker, Age-group and Gender Identification from Children's Speech", *Computer Speech & Language*, vol.50, no.7, pp. 141-156, 2018.

[21] A. Sandryhaila and J.M.F. Moura, "Big Data Analysis with signal Processing on Graphs: Representation and Processing of Massive data sets with irregular structure," *IEEE Signal Processing Magazine*, vol.31, no.5, pp.80-90, 2014.

[22] S. Meignen, D.H Pham, and M.A. Colominas, "On the Use of Short-Time Fourier Transform and Synchrosqueezing-Based Demodulation for the Retrieval of the Modes of Multicomponent Signals," *Science Direct*, vol.178,no.01, pp.107760, 2021.

[23] W. Mustikarini, R. Hidayat, and A. Bejo, "Real-Time Indonesian Language Speech Recognition with MFCC Algorithms and Python - Based SVM," *IJITEE*, vol. 3, no. 2, pp. 55-67, 2019.

[24] A.H.Nour-Eldin, "Mel-Frequency Cepstral Coefficient-Based Bandwidth Extension of Narrowband Speech," *Interspeech 2008 ISCA*, vol. 2008, no. 08, pp. 1-12, 2008.

[25] T. Furoh, F. Takahiro, M. Nakayama, and n. Takanobu, "A study of degraded-speech identification based on spectral centroid," *I-INCE*, vol. 2014, no. 01, pp. 1-4, 2014.

[26] Hua-Peng Zhang, "On the Construction of Fuzzy Betweenness Relations from Metrics," *Science Direct, Fuzzy Set and System* vol. 2020, no. 390, pp. 118-137, 2020.

[27] J.Y. Choi, "Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing," Psychonomic *Society*, vol. 2018, no. 2, pp. 1-12, 2018.

[28] C. Gowrishankar, "Properties of Composition of Fuzzy Relations and its Verifications," *International Journal of Management and Humanities (IJMH)*, vol. 4, no. 1, pp. 1-5, 2020.

[29] J.H.L. Hansen and H. Boril, "On the Issues of Intra-Speaker Variability and Realism in Speech, Speaker, and Language Recognition Tasks," *Science Direct*, vol. 5, no. 16, pp. 1-51, 2018.