Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 249–254

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Artificial Intelligence Methods for Smart Cities (AISC 2021)
November 1-4, 2021, Leuven, Belgium

Demographic Fairness in Multimodal Biometrics: A Comparative Analysis on Audio-Visual Speaker Recognition Systems

Gianni Fenu[a], Mirko Marras[b,*]

[a] *University of Cagliari, Cagliari, Italy*
[b] *EPFL, Lausanne, Switzerland*

Abstract

In urban scenarios, biometric recognition technologies are being increasingly adopted to empower citizens with a secure and usable access to personalized services. Given the challenging environmental scenarios, combining evidence from multiple biometrics at a certain step of the recognition pipeline has been often proved to increase the performance of the biometric-enabled recognition system. Despite the increasing accuracy achieved so far, it still remains under-explored how the adopted biometric fusion policy impacts on the quality of the decisions made by the biometric system, depending on the demographic characteristics of the citizen under consideration. In this paper, we investigate the extent to which state-of-the-art multimodal recognition systems based on facial and vocal biometrics are susceptible to unfairness towards legally-protected groups of individuals, characterized by a common sensitive attribute. Specifically, we present a comparative analysis of the performance across groups for two deep learning architectures tailored for facial and vocal recognition, under seven fusion policies that cover different pipeline steps (feature, model, score and decision). Experiments show that, compared to the unimodal systems alone and the other fusion policies, the multimodal system obtained via a fusion at the model step leads to the highest overall accuracy and the lowest disparity across groups.

Keywords: Multimodal; Biometrics; Biometric Authentication; Face Recognition; Voice Recognition; Fairness; Bias; Discrimination.

1. Introduction

The attention received by biometric-enabled systems is increasingly growing according to several market research reports [1]. Biometrics are being integrated by institutions, companies, and organizations for access control, national identity management, forensics analysis, student's identity control, and so on. These increasing operational and security demands are changing biometrics by shifting the focus from unimodal to multimodal biometrics [2]. Multimodal systems are considered more suitable because of weaknesses of unimodal systems in terms of reliability, e.g., considering another biometric in case the first one is not available, security, e.g., considering multiple ways to check

* Corresponding author. Tel.: +41-21-693-67-77; E-mail address: mirko.marras@acm.org

1877-0509
10.1016/j.procs.2021.12.236

the identity of a user, and coverage, e.g., ensuring authentication for very large populations. Multimodal systems do, however, add some complexity on how the sample input and the templates are used to provide a single authentication or identification decision, giving raise to a wide range of novel biometric fusion policies still under active research.

Meanwhile, recent research has shed light on issues regarding the fairness of biometric technologies, that may end up being demographically biased [3]. Indeed, certain groups of individuals (e.g., based on gender or ethnicity) may therefore suffer from statistically different outcomes in case the system is biased. Preventing that these systems emphasize demographic unfairness is considered a fundamental objective, often regulated by laws [4]. Considering a biometric recognition scenario, this biased behavior means that false positive and/or false negative error rates may systematically vary across demographic groups [5, 6]. The majority of works on fairness issues in biometric systems have been focused on facial [7, 8, 9] and vocal recognition [10, 11]. While preliminary studies do exist [12], the area of multimodal systems has by no means been researched exhaustively yet and still remains under-explored.

The aim of this work is to analyze multimodal systems using a discrimination-aware perspective. To this end, we formalize a fairness definition that serves as a ground for the assessment of disparate outcomes across demographic groups. We subsequently study the extent to which multimodal recognition systems based on facial and vocal biometrics emphasize group discrimination according to the considered fairness definition, depending on the adopted biometric fusion policy. The main contributions are therefore (i) a general formulation of algorithmic discrimination for multimodal systems, applied in this work in the context of audio-visual recognition, and (ii) a comprehensive analysis of causes and effects of biased outcomes including a fusion performance analysis on three public datasets according to gender-based demographic groups. Experiments show that individuals belonging to a certain demographic group systematically experience higher error rates, but the difference in error rates achieved by the multimodal system is smaller than the difference achieved by unimodal systems alone. Specifically, while being the most accurate, multimodal systems that perform a fusion at model step show also the highest fairness among the considered settings.

2. Experimental Methodology

2.1. Fairness Definition

Discrimination is defined as treating in a systematically different way a person or particular group of people characterized by a protected sensitive attribute (e.g., a group identified by the members' gender or ethnicity). For the purpose of studying discrimination in multimodal systems, we now formulate mathematically *algorithmic discrimination* for this class of systems by extending the definition adopted for unimodal biometric systems in [6]. For the sake of clarity, in our work, we focus on authentication tasks, leaving the case of identification tasks as a natural outlet for future work. First, we assume that a multimodal system is composed by $\mathcal{B}$ biometrics, e.g., $\mathcal{B} = \{Face, Voice\}$ (the biometrics used in this work). For each biometric b, we consider a machine-learning model $\mathcal{M}_b : (\mathcal{I}, \mathcal{I}) \rightarrow \{0, 1\}$ that receives a pair of input biometric examples from a set of individuals $\mathcal{I}$ and is optimized to correctly predict whether those biometric examples come from the same individual (0: different; 1: same). We also assume that $\mathcal{M}^p_\mathcal{B} : \mathcal{M}^* \times (\mathcal{I}, \mathcal{I}) \rightarrow \{0, 1\}$ is a multimodal system that receives a set of machine-learning models (one for each biometric $b \in \mathcal{B}$) and a pair of input biometric examples (zero-shot authentication scenario), and applies a fusion policy p (e.g., a fusion of the matching scores or of the decisions taken by the machine-learning models) to derive the combined authentication decision. We assume that there is a goodness criterion in the authentication task, given trial pairs $\mathcal{D} = \{(s_i, t_i)\}$ with $i = 1, \cdots, |\mathcal{D}|$:

$$\mathop{\mathbb{E}}_{(s_i,t_i)\,\in\,\mathcal{D},\,\mathcal{M}^p_\mathcal{B}} \begin{cases} M^p_\mathcal{B}(s_i, t_i) & \mathcal{I}_{s_i} = \mathcal{I}_{t_i} \\ 1 - M^p_\mathcal{B}(s_i, t_i) & \mathcal{I}_{s_i} \neq \mathcal{I}_{t_i} \end{cases} \tag{1}$$

where $\mathcal{I}_{s_i}$ and $\mathcal{I}_{t_i}$ are the individuals the two biometric examples s_i and t_i belong to. In other words, the goal is to maximize the cases where the multimodal system (i) gives access to the biometric-enabled service when both the examples come from the same individual, i.e., $M^p_\mathcal{B} = 1$ when $\mathcal{I}_{s_i} = \mathcal{I}_{t_i}$, and (ii) does not allow to access otherwise, i.e., $M^p_\mathcal{B} = 0$ when $\mathcal{I}_{s_i} \neq \mathcal{I}_{t_i}$. The $\mathcal{I}$ individuals can be characterized according to C sensitive attributes, each defined as C_d, with $d = 1, ..., |C|$, e.g., $C_1 = Gender = \{Male, Female\}$ (the sensitive attribute *Gender* has two classes in this example; we acknowledge that the gender is by no means a binary construct, and what we consider is a binary feature, as current public datasets offer.). The particular class $k = 1, ..., \mathcal{K}$ for a given sensitive attribute d and a given example is noted as $C_d(s_i)$, e.g., $C_1(s_i) = Male$. We assume that the number of examples for each class in $\mathcal{D}$ is significant.

We assume that a multimodal system $M^p_{\mathcal{B}}$ discriminates the group represented with class k of the sensitive attribute d (e.g., Gender) while authenticating, in case the performance achieved on the subset of authentication trial pairs where $C_d(s_i) = k$ is significantly lower than what is achieved on any another class of the sensitive attribute d.

2.2. Datasets

Our discrimination-aware analysis was conducted on multimodal systems based on face and voice biometrics, considering demographic groups based on the gender. We used a large audio-visual dataset to train the unimodal systems and three audio-visual datasets coming from diverse contexts to test. The datasets are described below.

- **Training Dataset**. *VoxCeleb1-Dev* [13] is an audio-visual speaker identification and authentication dataset collected from Youtube, including 21,819 videos from 1,211 identities (61% male). This dataset represents the only one large enough for training deep biometric models, due to the wide range of users and samples per user.
- **Testing Dataset #1**. *VoxCeleb1-Test* [13] is an audio-visual speaker identification and authentication dataset collected by from Youtube, embracing 677 videos from 40 identities.
- **Testing Dataset #2**. *MOBIO* [14] is a face and speaker recognition dataset collected by from laptops and mobile phones under a controlled scenario, including 28,800 videos from 150 identities.
- **Testing Dataset #3**. *AveRobot* [15] is an audio-visual biometric recognition dataset collected under robot assistance scenarios, including 2,664 videos from 111 identities.

2.3. Unimodal Systems

Our study is focused on combining facial and vocal biometrics. Before describing the fusion policies, we provide an overview of the characteristics and the operations performed by each individual unimodal system, as follows.

Face Model. Let $A_f \subset \mathbb{R}^{m \times n \times 3}$ denote the domain of RGB images with $m \times n \times 3$ size. Each input image $a_f \in A_f$ was pre-processed in order to detect the bounding box and key points (two eyes, nose and two mouth corners) of the face. The affine transformation was used to align the face. The image was then resized and each pixel value was normalised in the range [0,1]. The resulting intermediate facial image, defined as $S_f \subset \mathbb{R}^{m \times n \times 3}$, was provided as an input to a deep neural network. This network performed an explicit feature extraction which produced fixed-length embeddings in $D_f \subset \mathbb{R}^e$. We denote such a stage as $\mathcal{D}_{f\theta_f} : A_f \rightarrow D_f$. This embedding was used to make the decision.

Voice Model. Let $A_v \subset \mathbb{R}^*$ denote the domain of waveforms digitally represented by an intermediate visual acoustic representation $S_v \subset \mathbb{R}^{k \times *}$, such as a spectrogram or a filter-bank. Each audio $a_v \in A_v$ was converted to single-channel. The spectrogram was then generated in a sliding window fashion using a Hamming window, generating an acoustic representation s_v that corresponds to the audio a_v. The resulting representation was provided as an input to a deep neural network. This network performed an explicit feature extraction which produced fixed-length embeddings in $D_v \subset \mathbb{R}^e$. We denote such a stage as $\mathcal{D}_{v\theta_v} : S_v \rightarrow D_v$. This embedding was used to make the decision.

ResNet-50, known for good performance on visual and acoustic modalities [13], was used as a backbone for both models. The input layers of the original architecture were adapted to the corresponding biometric, and the fully-connected layer at the top of the original architecture was replaced by a flatten layer and a fully-connected layer whose output was an embedding of size 512. The face model was fed with images of shape (112, 112, 3), while the voice model received spectrograms of shape (512, 300, 1). More details on the implementations can be found in [16].

2.4. Fusion Policies and Recognition Protocol

The pre-trained models that represent the unimodal systems were merged under seven different fusion policies. These policies are well-known to achieve state-of-the-art performance in several multimodal scenarios [2] and cover different steps of the pipeline, from the feature step till the decision step. For each authentication trial ($s_i = (s_{f_i}, s_{v_i}), t_i = (t_{f_i}, t_{v_i})$), where f indicates face examples and v indicates voice examples, the fusion was performed as follows:

[Face | Voice]_Only. The embedding vectors of the two facial (vocal) examples s_{f_i} and t_{f_i} (s_{v_i} and t_{v_i}) were extracted by the face (voice) model, and an authentication decision was made (1 if the similarity score between the vectors was greater than a predefined threshold and so biometric examples came from the same individual, 0 otherwise).

Feature. The embedding vectors of the biometric examples s_{f_i} and s_{v_i} were extracted by the respective model and concatenated. The same process was repeated with the biometric examples t_{f_i} and t_{v_i}. The authentication decision was finally made (1 if the similarity score between the two concatenated vectors was greater than a predefined threshold so biometric examples came from the same individual, 0 otherwise).

Model. The two unimodal models were fused: the embedding layers of the two unimodal models were concatenated, and the concatenation layer was connected to a dense neural network. The fused model was then fine-tuned under a supervised classification task, by receiving face-voice pairs as inputs. The same experimental setting and hyper-parameters adopted in [16] were considered. The embedding vectors of the biometric examples s_{f_i} and s_{v_i} were extracted from the fine-tuned unimodal models and concatenated. The same process was repeated with the biometric examples t_{f_i} and t_{v_i}. The authentication decision was made (1 if the similarity score between the two concatenated vectors was greater than a predefined threshold so biometric examples came from the same individual, 0 otherwise).

Score_[Avg | Max | Min]. The embedding vectors of the two face examples s_{f_i} and t_{f_i} were extracted by the face model and the similarity score between the vectors was calculated. The same process was repeated with voice examples s_{v_i} and t_{v_i}. The authentication decision was made (1 if the *mean / max / min* between the face and the voice similarity scores was greater than a predefined threshold so biometric examples came from the same individual, 0 otherwise).

Decision_[AND | OR]. The embedding vectors of the two face examples s_{f_i} and t_{f_i} were extracted by the face model, and an authentication decision was made (1 if the similarity score between the vectors was greater than a predefined threshold and so both face examples came from the same individual, 0 otherwise). The same process was repeated with the voice examples s_{v_i} and t_{v_i}. The final decisions was the logical *AND / OR* performed on the unimodal decisions.

The following evaluation protocol was considered for each fusion policy. For each dataset, 30 users were randomly selected (15 males; 15 females) and, for each user, 100 positive (both examples from the same individual) and 100 negative (examples from two different individuals) authentication trials were randomly created (6,000 trials in total for each fusion policy). Then, the EER[1] was computed to assess the recognition performance (the lower it is, the higher the performance). To ensure a fair comparison across policies[2], the threshold was optimally chosen for each model on that policy by considering the EER security level of that model under that policy (e.g., for the model-step policy, the similarity scores between embedding vectors extracted from the fine-tuned unimodal models for all authentication trials were obtained, and the similarity threshold that led to the same proportion of false accepts and false rejects along trials for that policy, i.e., the EER, was chosen). Hence, thresholds were different across policies.

3. Experimental Results

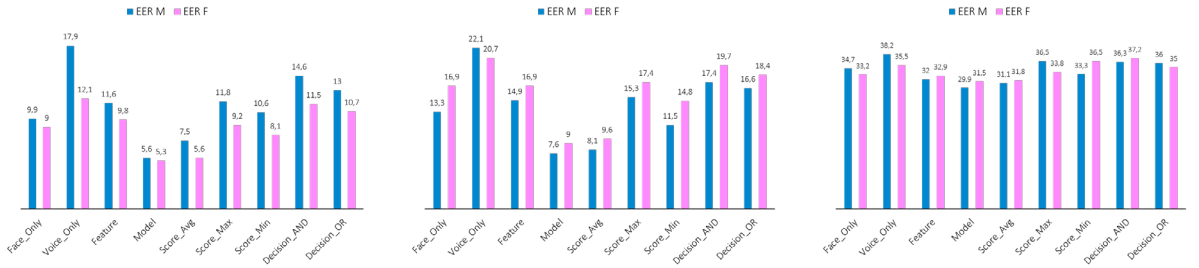
In this section, we empirically evaluate the unimodal systems separately and the multimodal systems obtained by means of the considered fusion policies in terms of accuracy and fairness. We aim to answer two research questions:

- RQ1. Does the policy adopted to fuse facial and vocal biometrics impact on accuracy and fairness of the system?
- RQ2. Does the best performing multimodal system result in the same error patterns across demographic groups?

3.1. RQ1 Fusion Policy Comparison

Figure 1 reports the EER for each demographic group on the three considered testing datasets. It can be observed that the lowest EER was achieved by performing a model-step fusion on all the datasets, in turn, leading to the lowest EER for both male and female individuals. Score-step fusion policies also showed a good recognition performance, while the other policies led to performance close or even worse than the facial unimodal system only. In VoxCeleb1-Test, female individuals obtained an EER always lower than that of the male individuals, for any type of fusion considered. Conversely, the other two datasets resulted in lower EERs for male individuals. In terms of disparity in EER between genders, the best-performing fusion policy is often also the fairest one (the one with the lowest disparity), both considering the unimodal systems and the other multimodal systems. For instance, in VoxCeleb1-Test, the disparity is of 0.9 for the face unimodal system and of 5.8 for the voice unimodal system, while decreases

[1] The Equal Error Rate (EER) indicates the point where the proportion of false accepts (FAR) is equal to the proportion of false rejects (FRR).
[2] The use of a threshold to determine the positive/negative authentication decision is internally present for all fusion policies.



(a) VoxCeleb1-Test (b) MOBIO (c) AveRobot

Fig. 1: Equal error rates for gender-based groups across three public datasets.

to 0.3 under the model-step multimodal system. Similar observations were confirmed also in the other two datasets. In MOBIO (and AveRobot), the disparity is of 3.6 (1.5) for the face unimodal system and of 1.4 (2.7) for the voice unimodal system, while reaches 1.4 (1.5) under the model-step multimodal system, being also more accurate overall. The disparity under the model-step multimodal system is often not higher than that of the unimodal systems, but the EER achieved by the multimodal system is significantly lower than that of unimodal systems, thus meeting a better trade-off between accuracy and fairness. In contrast to several fairness studies in other systems enabled by machine learning (e.g., recommendation), our findings show that the most accurate system is often also the fairest one.

3.2. RQ2 Error Analysis

Table 1 shows the truth and the predicted decisions made by unimodal systems and the multimodal system adopting model-step fusion on the VoxCeleb1-Test testing dataset. The *Target* column indicates the ground-truth decision to make (0 = different user; 1 = same user). The *Face_Only*, *Voice_Only*, and *Model* columns report the predicted decision returned by the respective system. The *Total Cases* column reports the number of trials resulted in that truth-predicted decision outcome. The remaining three columns report the number of those total cases that are associated to a male-male comparison (M-M: the examples of the pair both come from male individuals), a female-female comparison (F-F: the examples of the pair both come from female individuals) or a inter-gender comparison (M-F or F-M: the examples of the pair come from different genders). The top half part reports the cases where the multimodal system made an error, while the bottom half part refers to cases where the multimodal system made a correct decision (i.e., the values in the *Target* and the *Model* columns are equal). For instance, the first row shows the cases where both unimodal systems were right while the multimodal system was wrong (66 cases). While being recognized as reasonably fair (EER M = 5.6; EER F = 5.3), the considered multimodal system made errors differently with respect to the unimodal systems, depending on the gender of the individual under consideration. When the multimodal system made a wrong prediction (the top part), more errors were made for male (female) individuals when the face unimodal system, i.e., the most accurate unimodal system, was right (wrong). Therefore, the most accurate unimodal system has the highest influence in the fairness of the resulting multimodal system. Conversely, when the multimodal system made a correct prediction (the bottom part), more errors were made for male (female) individuals when the two unimodal system disagree (agree). This means that, while overall reaching a lower disparity in errors between the genders, the multimodal system made different types of errors across genders, based on the outcome of the unimodal systems.

4. Conclusions and Future Works

In this paper, we presented a comprehensive analysis of multimodal recognition systems based on facial and vocal biometrics, according to a new discrimination-aware perspective. The results showed that the two unimodal systems are highly biased across demographic groups. In particular, we often observed larger performance differences in voice recognition than face recognition across demographic groups based on the gender. These performance disparities are reduced when the unimodal systems are fused, while simultaneously achieving an overall lower error rate. This means that the benefit of fusing biometrics may often be twofold, i.e., on accuracy and fairness. We also revealed different

Target	Fusion Policy			# Comparisons			
	Face_Only	*Voice_Only*	*Model*	*Total Cases*	*M-M*	*F-F*	*M-F/F-M*
0	0	0	1	66	**40**	26	0
1	1	1	0				
0	1	0	1	95	38	**55**	2
1	0	1	0				
0	0	1	1	72	**43**	27	2
1	1	0	0				
0	1	1	1	116	51	**65**	0
1	0	0	0				
0	1	1	0	100	41	**59**	0
1	0	0	1				
0	0	1	0	743	**424**	291	28
1	1	0	1				
0	1	0	0	324	**175**	125	24
1	0	1	1				
0	0	0	0	4484	1438	**1602**	1444
1	1	1	1				
Total				6000	2250	2250	1500

Table 1: Comparison of truth-predicted decisions made by the unimodal systems and the multimodal system adopting model-step fusion.

error patterns for different demographic groups. In next steps, we will investigate how to incorporate user-specific algorithmic discrimination, going beyond group-based fairness. Additionally, the analysis of other covariates, such as the age, biometrics, such as fingerprints and iris, and tasks, such as best sample selection via quality assessment [17] instead of zero-shot scenarios, will be conducted. Other future directions include the development of new methods to detect bias in the fusion process and new fusion policies that meet better accuracy-fairness trade-offs.

References

[1] D. Thakkar, "Global biometric market analysis: trends and future prospects," 2018. Accessed: July 18, 2021. [Online]. Available: https://www.bayometric.com/global-biometric-market-analysis/.

[2] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187–205, 2019.

[3] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[4] B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making: "Right to Explanation"," *AI Magazine*, 2017.

[5] C. Rathgeb, P. Drozdowski, N. Damer, D. C. Frings, and C. Busch, "Demographic fairness in biometric systems: What do the experts say?," *arXiv preprint arXiv:2105.14844*, 2021.

[6] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "Sensitivenets: Learning agnostic representations with application to face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2158–2164, 2020.

[7] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE transactions on biometrics, behavior, and identity science*, vol. 3, no. 1, pp. 101–111, 2020.

[8] I. Serna, A. Peña, A. Morales, and J. Fierrez, "Insidebias: Measuring bias in deep networks and application to face gender biometrics," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3720–3727, IEEE, 2021.

[9] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Post-comparison mitigation of demographic bias in face recognition using fair score normalization," *Pattern Recognition Letters*, vol. 140, pp. 332–338, 2020.

[10] G. Fenu, H. Lafhouli, and M. Marras, "Exploring algorithmic fairness in deep speaker verification," in *Proc. of the International Conference eon Computational Science and Its Applications (ICCSA)*, pp. 77–93, 2020.

[11] G. Fenu, G. Medda, M. Marras, and G. Meloni, "Improving fairness in speaker recognition," in *Prof. of the European Symposium on Software Engineering (ESSE)*, p. 129–136, ACM, 2020.

[12] M. Köppen, A. Soria-Frisch, and J. Acedo, "Fairness-based parameter selection in multi-modal biometric authentication," in *2012 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 979–985, IEEE, 2012.

[13] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[14] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernockỳ, N. Poh, J. Kittler, A. Larcher, C. Levy, *et al.*, "Bi-modal person recognition on a mobile phone: using mobile phone data," in *IEEE Inter. Conf. on Multimedia and Expo Workshops*, pp. 635–640, IEEE, 2012.

[15] M. Marras, P. A. Marín-Reyes, J. Lorenzo-Navarro, M. C. Santana, and G. Fenu, "Averobot: An audio-visual dataset for people re-identification and verification in human-robot interaction.," in *ICPRAM*, pp. 255–265, 2019.

[16] M. Marras, P. A. Marín-Reyes, J. Lorenzo-Navarro, M. Castrillón-Santana, and G. Fenu, "Deep multi-biometric fusion for audio-visual user re-identification and verification," in *International Conference on Pattern Recognition Applications and Methods*, pp. 136–157, Springer, 2019.

[17] D. Freire-Obregón, K. Rosales-Santana, P. A. Marín-Reyes, A. Penate-Sanchez, J. Lorenzo-Navarro, and M. Castrillón-Santana, "Improving user verification in human-robot interaction from audio or image inputs through sample quality assessment," *Pattern Recognition Letters*, vol. 149, pp. 179–184, 2021.