NaijaFaceVoice: A Large-Scale Deep Learning Model and Database of Nigerian Faces and Voices

Publisher: IEEE

Cite This

PDF

Adekunle Anthony Akinrinmade; Emmanuel Adetiba; Joke A. Badejo; Oluwadamilola Oshin

All Authors

420

Full

Text Views

Open Access

Comment(s)

-
-
-
-

-

Abstract

Document Sections

Show Full Outline
Authors
Figures
References
Keywords
Metrics

Abstract:

The fusion of two or more traits in multimodal biometrics generally improves recognition accuracy. The question is, by how much? Large-scale databases are better suited for training deep learning models for better generalization and accuracy. Therefore, a large-scale multimodal database is beneficial. However, publicly available large-scale multimodal databases are scarce, especially for faces and voices. Again, because a face image is 2-D while a voice is 1-D, there is the challenge of the best way to fuse both. Therefore, improvements owing to fusion have hitherto yielded marginal improvements. This study proposes a semi-automated curation algorithm for the extraction of the faces and voices of target individuals in videos to create a large-scale face-voice database. The curation technique involves observing the positions at the time of the occurrence of the target subject’s faces and voices in videos. These positions are supplied to a MATLAB2017b script that detects the faces in the observed regions, crops, resizes, auto-labels, and writes them to the disk. A second MATLAB2017b script, extracts the audio content within the observed regions, auto-labels, and writes the voice segments to the disk. The created database named NaijaFaceVoice consists of 2,656 subjects with over 2 million faces and 195 hours of utterances. The database was employed to develop a large-scale recognition system that leveraged Convolutional Neural Networks. Robust fusion methods incorporating the proposed Spectrogram-Voting concept significantly improved performance achieving a record equal error rate of 0.0003519%, an improvement by a factor of over 450.

The database creation pipeline and recognition models.

Published in: IEEE Access (Volume: 11)
Page(s): 58228 - 58243
Date of Publication: 05 June 2023
Electronic ISSN: 2169-3536
DOI: 10.1109/ACCESS.2023.3282618
Publisher: IEEE
Funding Agency:

SECTION I.

Introduction

Existing bimodal face-voice databases are scarce, the ones available are small-scale (with a small number of subjects and samples per subject) [1], [2], [3], [4]. Therefore, research results using such databases are not generalizable because of the limited number of subjects, samples per subject, and demographic diversity [1]. Although automated methods have been employed to create large-scale unimodal databases such as face and voice, there is the problem of the unavailability of a large-scale bimodal face-voice database required for developing recognition models based on the fusion of both traits. As a result of this limitation, most research works in the face-voice field have focused on the small-scale domain while other researchers circumvent the problem by creating virtual face-voice databases from stand-alone face and voice databases of different individuals, making assumptions that the faces and voices used belonged to the same individuals [5]. The shortcoming of this approach is that it only mimics real-life scenarios. Although the large-scale VoxCeleb speaker recognition database was created using the VGGFace database, the number of unique individuals in both databases differed. In addition, there is currently no one-to-one mapping of individuals with the facial images in the VGGFace database with those having the corresponding voice segments in the VoxCeleb database. This situation is the same as that of the VGGFace2 and VoxCeleb2 pair. Consequently, these databases are suitable for unimodal recognition research. Therefore, this research focuses on creating a large-scale bimodal database of faces and their corresponding voice samples to facilitate diversity in biometric research, especially for the black population [6].

The first goal of this study is to create a large-scale database of faces and corresponding voice samples using a semi-automated curation pipeline that can be used for face, voice, or face-voice recognition. The database is annotated to make it relevant for gender recognition based on either a combination of modalities as well as language recognition. The second goal is to develop a robust fusion technique to significantly improve the recognition performance compared to the best result in unimodal cases. Three problems identified in the literature must be addressed to achieve these goals. First, although there have been improvements due to the fusion of these traits, these have been marginal and most studies have been conducted on small-scale databases. Second, there is a lack of a large-scale bimodal database of face and voice samples with a one-to-one mapping of the modalities for the database subjects needed for evaluation. This implies the need to create such a database. The third reason is that the existing methods of large-scale database creation focused on the creation of either a large-scale face or a large-scale voice database but not both. Well-known state-of-the-art databases created using automated methods such as MegaFace, MSCeleb-1M, CACD, and VoxCeleb2 have been found to contain errors [7], [8], [9]. A semi-automated approach to curation was thus employed in this study to mitigate errors. The manual approach ensures correctness before auto-curation which is followed by a final manual cross-check. The semi-automated curation algorithm extracts the faces and voices of target individuals in videos leveraging YouTube, a rich video source. The faces and voices extracted from the videos were used to create a large-scale face-voice database. The created database was then divided into two non-overlapping partitions in the ratio of 80:20 for training and testing the developed recognition systems. The main contributions of this study are as follows:

i. introduced the concept of Spectrogram-Voting and Vote-Code generation, these methods are new in the literature regarding speaker recognition, and outperforms speaker recognition results in the literature on the state-of-the-art VoxCeleb as detailed in the experimental results of Table 9, 12, and 13,

ii. developed a semi-automated pipeline for the curation of faces and corresponding voice samples of individuals in videos that is re-usable, this method is unique to this paper and unlike other pipelines that generate either a face or a voice database, this method generated both, to address the scarcity of large scale bimodal database of faces and corresponding voices, the experimental results justifying the applicability of these databases for recognition research are captured in Tables 6–7,

iii. created a new bimodal large-scale database of Nigerian faces and their corresponding voice samples (in different languages) for 2656 subjects with over 2 million face samples and about 150,000 voice samples, the details of the database are contained in Tables 1–4,

iv. developed a robust fusion method for face and voice traits that significantly improved verification performance by rapidly reducing the EER compared with state-of-the-art methods as detailed in the experimental results in Table 14.

TABLE 1 Summary of NaijaFaceVoice Statistics

TABLE 2 NaijaFaceVoice Distribution by Utterance Length

TABLE 3 NaijaFaceVoice Database Distribution by Gender

TABLE 4 NaijaFaceVoice Distribution by Language

TABLE 5 Structure of the CNN Architecture Used for Classification

TABLE 6 The Relative Purity of the NaijaFace Database

TABLE 7 The Relative Purity of the NaijaVoice Database

TABLE 8 Performance Evaluation of Unimodal Face

TABLE 9 Performance Evaluation of Unimodal Voice Recognition Systeme

TABLE 10 Performance of Feature-Level Fusion

TABLE 11 Performance of Score and Decision-Level Fusion

TABLE 12 Benchmarking Vote-Code Concept With State-of-the-Art

TABLE 13 Benchmarking Spectrogram-Voting Concept With State-of-the-Art

TABLE 14 The Improvement Due to Fusion

SECTION II.

Related Works

Previous methods of face database creation involved inviting candidates to a specific location for acquisition- This manual process is usually divided into sessions and requires cooperation from candidates hence it is tedious. In this process, the candidates are required to make different facial expressions and head rotations, and the illumination and distances of subjects from the camera are manually varied to create a close-to a real-life scenario. This method requires considerable cooperation from the candidates. An example is the creation of a small-scale Olivetti Research Laboratory (ORL) face database containing 10 images of 40 individuals each, which took about two years [10]. Another example is the Facial Recognition Technology (FERET) database [11] containing 14,126 face image samples that were created in 15 sessions lasting 3 years. The creation of speaker recognition databases followed a similar procedure requiring candidates to be physically present at the recording studio. These speakers are made to read certain sentences or utter certain combinations of words; these manual processes take considerable time and effort- For example, the creation of the YOHO voice corpus [12] spanned 200 sessions of audio recordings.

More recently, researchers have adopted automated or semi-automated methods that eliminate the constraints of individuals coming to specific locations for capturing and changing accessories during capture or reading several long sentences. These methods rely on extracting these traits from the internet- and the removal of these constraints has pioneered the creation of large-scale databases. One example of such a face database is the VGGFace [13] consisting of 2.6 million face images from 2,622 subjects, and the other is the VGGFace2 database [14] consisting of 3.31 million face images obtained from 9,131 subjects through Google Image Search. In the speaker recognition domain, the GBR-ENG database [15], which consists of 6,000 utterances from 600 subjects was created using utterances extracted from telephone conversations. VoxCeleb [16] and VoxCeleb2 [9] are both speaker recognition databases created using a fully automated pipeline; the former is a database of 153,516 utterances from 1,251 celebrities whereas the latter contains 1,128,246 utterances from 6,112 celebrities. These automated/semi-automated methods reduce errors associated with the manual acquisition process and database creation time, which would otherwise have taken several years.

Some examples of recent face recognition works were the use of a firefly optimization technique to reduce the dimension of the local ternary pattern (LTP) and binary robust invariant scalable key (BRISK) features classified using a deep belief network (DBN) on the AT&T and Yale databases. The method improved the recognition accuracy by up to 20%, however, the database employed was small-scale [17]. The objective of [18] was to improve face recognition by extracting more independent features using logarithmic independent component analysis (Log-ICA) before classification. The method was effective at recognizing faces in noisy scenarios and improved the accuracy by 10% on a small-scale Yale database [19]. Face recognition by the estimation of the participation of face pixels in identification using type-II fuzzy logic followed by K nearest neighbors (KNN) using Euclidean distance for classification has also been attempted [20]. The use of scale invariant feature transformation (SIFT) for both feature extraction and matching was adopted by [21]. However, although good performance was obtained, the databases used were small-scale. The use of bottleneck residual blocks and fast down sampling at the earlier layers of a light-weight convolutional neural network (CNN) with the addition of more feature maps at the later stages was explored by [22] on the Labeled Faces in the Wild (LFW) which is also a small-scale database. Although they attained a recognition performance of 99.73%, this dropped to 91.3% when applied to a large-scale MegaFace database. This result clearly shows that small-scale databases cannot be used for generalization.

In the speaker recognition space, [23] mitigated the duration mismatch using a refinement approach between training and inference. They modified parts of the deep neural network (DNN) parameters with full recordings to decrease the mismatch and generate embeddings for cosine distance scoring. Reference [24] investigated speaker recognition in a multi-speaker environment. Diarization was implemented based on x-vectors and agglomerative hierarchical clustering (AHC) at the front end with probabilistic linear discriminant analysis (PLDA) at the back end. Both teams of researchers achieved good results on the Speakers in the Wild (SITW) database which is also a small-scale database. As shown earlier, these results cannot be generalized. Reference [25] developed a loss function based on softmax cross-entropy with adaptive parameters that reduced the training time and improved accuracy. Reference [26] improved speaker recognition by analyzing models based on ResNet. They employed variants of ResNet by varying the loss functions trained using classification-based and metric learning objectives. The work by both research teams was on the large-scale VoxCeleb database, they obtained good results, however, these results could be improved if fusion was considered. The fusion of multiple systems was adopted in [27]. Embeddings were used as features from DNNs fine-tuned with additive angular margin loss with PLDA and cosine distance explored as back-ends. However, diminishing returns in performance were observed as the number of systems used in the fusion increased. In addition, a marginal improvement was realized from fusion because the same modality was used on a unimodal database.

In the face-voice recognition space, embeddings were extracted as features from both face and voice with classification performed using PLDA or CNN on NIST SRE 2018 and 2019 exploring feature-level fusion [28] and [29]. The improvement compared to the best unimodal scenario as a result of fusion was marginal. This was an improvement from the EER of 0.375% to 0.347% [28]. In [29], the EER improved from 1.66% to 1.11%. Additionally, these databases are not publicly available. Owing to the scarcity of bimodal face-voice databases, some researchers have had to create private databases for research. Examples of such studies include [30]. They employed EigenFace and principal component analysis (PCA) as facial features and cepstral coefficients as voice features to explore feature- and score-level fusion on a private dataset of 100 subjects. Reference [31] extracted HOG and LBP as facial features and MFCC for a voice which was combined at the feature level followed by classification using KNN on a private dataset of 27 candidates. Using EigenFace and PCA as facial features and linear prediction coefficients (LPC), linear prediction cepstral coefficients (LPCC), and Mel frequency cepstral coefficient (MFCC) voice features, [32] explored feature-level and score-level fusion and a Gaussian mixture model (GMM), artificial neural network (ANN), and support vector machine (SVM) as classifiers on a private database of 100 individuals. In [33], Eigenface, linear discriminant analysis (LDA), and Gabor filters were employed as face features. For audio features, MFCC, LPCC, and time domain statistics such as bass, baritone, tenor, alto, and soprano were considered for 100 subjects using GMM as the classifier while exploring feature- and score-level fusion. In these cases, the private databases used were not only small-scale but also not publicly available. This clearly shows the need to create such a large-scale bimodal face-voice database. In the cases mentioned above, the improved performance as a result of fusion compared to the best unimodal cases was marginal. Although [34] explored a hybrid fusion combining both feature-level and score-level fusion on a private dataset of 100 candidates, the improvement in performance as a result of fusion was still marginal; an improvement in EER from 1.373% to 0.64%, a reduction by a factor of 2.

However, these existing automated methods of large-scale database creation focus on unimodal databases, which use the information on the internet to produce either a large-scale face database or a large-scale voice database. There, there is a need for an algorithm that simultaneously extracts both modalities from online videos and ensures correlation in a single procedure to overcome this limitation. Similarly, large-scale recognition models are based on either face or voice modalities. Therefore, there is a need for the development of a model capable of performing recognition based on the combination of both traits on a large scale that will significantly improve performance because of fusion.

In [35], the face features employed were the Histogram of Oriented Gradients (HOG) and Local Binary patterns (LBP). These were transformed using the Linear Discriminant Analysis (LDA). The voice feature used was the Mel-Cepstral Frequency Coefficients (MFCC) modeled using Gaussian Mixture Model (GMM) and the Universal Background Models (UBM). The fusion of both traits was done using the Dempster–Shafer Theory. Similarly, [36], employed the LBP consisting of textural operators extracted as the face feature. The MFCC and median filter were used for the voice features. Fusion was achieved using a KNN classifier for both modalities. Other works employing the fusion of LBP and MFCC includes [37], where the face detection was based on Haar and AdaBoost and the feature extracted using LBP, a distance metric used to obtain the face matching score. MFCC features modeled by GMM, and maximum posterior probability was used to obtain the matching score for the voice trait. The fusion of both features was then done at the score level before authentication. Features such as LBP and Log Gabor have also been extracted from the face and the Linear Predictive Coding (LPC) and MFCC features extracted for the voice modality. There were four different fusion methods explored, these were; feature-level fusion by concatenation, score-level fusion applying mthe aximum mode technique, rank-level fusion usithe ng Borda count method, and decision-level fusion using the logical AND operator with KNN as the classifier [38].

In the work of [39], the face image served as the input to the visual modality leg of a model that generates a feature vector for the face. A spectrogram served as the input to the audio of the model to generate the feature vector for the voice. The fusion was a concatenation of the sparse feature representation from the voice and face modalities with an additional learning step that was jointly learned from both modalities. This fusion yielded an enriched sparse representation. The researchers in [40], proposed a robust fusion method of audio and visual modalities using combined cross-attention by taking advantage of the inter-modal and intra-modal relationships of both features. Temporal Convolutional Networks (TCN) and ResNet were used to extract the audio and visual features, respectively. The attended weighted features were obtained by a mechanism that correlated the individual and combined features simultaneously, this served as input to linearly connected layers to predict emotions. Reference [41] introduced a method that shows the degree of uncertainty the individual face and voice modality contributed to the fusion to improve performance. The facial and audio features were extracted using the fully connected layers of CNN models. The method uses the softmax match loss function to jointly learn calibrated and ranked individual traits uncertainty measures to estimate the amount of information each unimodal feature contributes to fusion in order to enhance fusion with respconcerningiction of emotions. In [5], the fusion of face and voice modalities was done at the feature level, their experiment include using features extracted from the face using raw pixels, Principal Component Analysis (PCA) and Discrete,Cosine Transform (DCT). The features from the audio were obtained using Vector Quantization (VQ) and MFCC. The facial and audio features were organized into matrices which were combined at feature level exploring concatenation, merging, and element-wise multiplication. These features were classified using ANN and KNN.

Reference [42], following the encoding of the audio signals to spectrograms, a face sub-networ,k and a voice sub-network were used to extract128-dimensionall features from each modality through the fully connected layers of their sub-networks. The features were L2-normalized before the combination in a way that exploits the best complementary information in each of the modalities that enriched the fusion using clustering based on orthogonality constraints.

SECTION III.

Methodology

The overview of the method adopted in this work is summarized in Figure 1. It consists of the following sections; the data acquisition pipeline, the dataset, preprocessing, the CNN architecture used for the training samples, the generated model, the feature level fusion, score level fusion, and the decision level fusion designs. The data acquisition pipeline is a semi-automated algorithm consisting of 6 stages. It accepts YouTube videos, curates the faces and corresponding voice samples of the individuals therein, and populates a dataset, referred to as NaijaFaceVoice. This database was partitioned into two non-overlapping segments in the ratio of 80:20 for training and testing custom CNN architectures with the hold-out cross-validation scheme. Before the training and testing of the CNN architectures, the face and voice samples were preprocessed and the generated spectrograms were enhanced. The CNN architecture Section comprises of a custom Face CNN and a custom Voice CNN. These were trained using the face and voice samples, respectively, to obtain the Face and voice recognition model, as shown in the Model section. These models were used to extract facial and audio features which were fused and used to train a custom Face-Voice-CNN to realize a face-voice recognition model.

FIGURE 1.
The architecture of the recognition system.

Show All

The testing samples were used to evaluate the performance of the developed recognition systems. As seen in the feature-level fusion section, face recognition and voice recognition models were used to extract facial and audio features. These features were then concatenated and fed to the face-voice recognition model to make the final decision. The score-level fusion section shows how the scores of predictions from the face recognition model and the Spectrogram-Voting scores adapted using the majority voting rule were combined using weigthe hted arithmetic mean. The decision-level fusion section of the diaram, depicts how the decision from the face recognition model and those from the voice recognition model were combined using the weighted majority voting rule to make the final decision.

SECTION IV.

Data Acquisition Method

Due to the non-availability of methods to evaluate large-scale databases created from the wild using fully automated algorithms [43] and the pervasive nature of noise associated with the fully automated methods for large-scale database creation [7], [8], [9], [44], [45], [46], this work adopted a semi-automated approach. The semi-automated approach is state-of-the-art [7], [14], [46], [47]. It consists of 6 stages:

- *Stage 1:*

  The first stage in the semi-automated curation pipeline was the compilation of the names of the subjects to be used in populating the large-scale database. This list was obtained from the internet (YouTube), it includes politicians, government officials, musicians, comedians, sports personalities, and actors. The next task entails searching YouTube using each name and the word ‘interview’ to ensure that the target candidate in the video was speaking similarly to [16]. Apart from this list, other popular Nigerian TV stations on YouTube such as Channels TV, Nigeria Television Authority (NTA), and so on with presence on YouTube were listed to capture other Nigerians who appeared on such programs but were not necessarily popular. The process was used to compile a list of 2,656 candidates of interest.

- *Stage 2:*

  Where possible, this stage involves downloading the top 5 videos of each of the targets using the list in stage 1.

- *Stage 3:*

  The procedure in this stage entails sampling the downloaded videos for each individual; the final video selection favored where the target spoke for the most prolonged duration or, where applicable, the video contained only the target subject.

- *Stage 4:*

  This involves the observation of the portions of the video where only the target's face appeared and the parts of the video where only the target was speaking. The start time is the time in the video when the target's trait (e.g. the face) appears in the video while the stop time is the time in the video when the face goes off-screen. These times are determined by the video being processed. Since a target's face or voice can appear in multiple parts of a video, the series of start and stop times are observed before curation. This is to ensure only the target's trait is curated and mitigate against errors. The series of pairs of start and stop instants in time were supplied to a curation algorithm that automatically extracts the information in the specified portions of the video.

- *Stage 5:*

  This is the automatic extraction of the faces and voice samples of the target from the video by a curation algorithm. The series of observations for the face in stage 4 is supplied to an automatic face curation algorithm developed using MATLAB 2017b which detects the faces in the supplied regions, and crops,kernel-based resizes to 80×80 pixels, auto-labels, and copies the faces to the disk. Figure 2 depicts the activities involved in this process. Similarly, the series of observations for the target's voice is supplied to an automatic voice curation algorithm developed using MATLAB 2017b which extracts the audio at the specified locations, concatenates them, splits the combined signal into smaller audio segments of 5 seconds duration, and auto-labels them. Figure 3 captures the flowchart for the audio curating process.

- *Stage 6:*

  This involved the manual cross-check and cleanup of errors in the curated samples. Such errors include false face detections, the capture of a non-target's face due to possible errors in the series of start/stop positions supplied to the script or the capture of non-target audio due to the possible time lag between the audio and visual streams. The manual clean-up was done by examining the folder location containing the curated faces to ascertain that it contains only the instants of faces of the specified target. If the location contains only the target's faces, no clean-up is required, otherwise, the non-face images detected or the non-target's faces in the folders are deleted. Similarly, the folder containing the curated audio is examined, and each audio is replayed to verify it belongs only to the target. Audio files not belonging to the target were removed from the folders. This way the corpus was manually cleaned by human effort as in [7], [14], and [47].

FIGURE 2.
Flowchart for semi-automation of face extraction from videos.

Show All

FIGURE 3.
Flowchart for semi-automation of audio extraction from videos.

Show All

SECTION V.
Dataset Description

NaijaFaceVoice contains more than 140,000 utterances totaling 195 hours of talk time and over 2 million face images from 2,656 Nigerian subjects. The male candidates represent 62.8% while the remaining 37.2% account for females. It is essentially a voice database (NaijaVoice) plus a face database (NaijaFace) with the positions of the traits in both databases corresponding to the appropriate candidate. The database is divided into 7 categories based on utterance length and annotated for language and gender. All speakers in the database are Nigerians from different ethnicities and the spoken languages span English, Hausa, Igbo, and Yoruba. The faces and voices were extracted from YouTube using a semi-automated pipeline, as a result, the faces in the database are characterized by varying degrees of blur, rotation, scale, translation, occlusion, illumination, and pose. The voice counterpart is characterized by background and channel noise, overlapping speech, and reverberation. Tables 1, 2, 3 and shows the summary statistics, and the distribution by utterance length, gender, and language respectively.

SECTION VI.

Unimodal Systems Design

Unimodal systems make use of a single trait for recognition. In this section, the research first explores the use of only the face trait for recognition. The study then proceeds to use only the voice trait for recognition. These recognition systems were implemented using custom CNNs with the structure captured in Table 5. CNN was considered the tool for the development of the model in this work because it has been proven to yield better performances compared with hand-crafted methods especially large-scale data samples [48]. Finally, to improve the performance of the voice recognition system, the concept of voting was adopted for the predictions of spectrograms generated from an utterance. The predictions were performed using the speaker recognition model obtained during the training of the CNN for speaker recognition. These predictions (votes) are counted and the candidate with the majority vote is deemed the owner of the utterance. This technique is referred to in this research as Spectrogram-Voting. Category 7 of NaijaFaceVoice with the statistics in Table 2 having the highest number of unique individuals and samples for the subjects was used for the experiment since CNNs rely on large sample sizes to better generalize predictions. This category consists of 1,661 subjects, each having at least 300 face samples and at least 25 voice samples.

A. Custom CNN-Based Face Recognition

The custom Face-CNN used in this research has the same architecture shown in Table 5 with a learning rate of 0.01 and an epoch of 4. The study randomly selected 300 face samples for each candidate. The face images were resized to 128×128 pixels and converted to grayscale. The ratio of the training to testing samples is 80:20.

B. Custom CNN-Based Speaker Recognition

The custom Voice-CNN made use of the architecture in Table 5. The only difference is the number of epochs and the learning rate used for the training set to 6 and 0.002, respectively. For each of the speakers, 25 voice samples were randomly selected. Spectrograms were generated using voice lengths of 0.4 seconds, a window size of 6 ms, 80% overlap, 512 Discrete Fourier Transform (DFT) points, and a pre-emphasis factor of 0.99 without regard for Voice Activity Detection (VAD). This resulted in a total of 300 spectrograms for each candidate. These spectrograms were converted to grayscale and resized to 128×128 pixels using image processing techniques then randomly divided in the ratio 80:20 for training and testing the custom Voice-CNN, respectively.

C. Spectrogram-Voting-Based Speaker Recognition

Speaker recognition falls into two parts, these are speaker identification and speaker verification. In speaker identification, the task is to identify who, in a closed set of speakers an utterance belongs to while speaker verification entails ascertaining if a speaker is whom he or she claims to be. The Spectrogram-Voting concept applies to both scenarios and is discussed in more detail in the sub-sections that follow

1) Speaker Identification Using Spectrogram-Voting

The Spectrogram-Voting method aims to improve the custom CNN-based speaker recognition result. The concept behind Spectrogram-Voting is to perform speaker identification at the utterance level rather than at the spectrogram level. This approach divides utterances in the dataset into a train/test ratio similar to the procedure for custom use of CNN, that is, the training to test utterance samples was in the proportion of 20:5. The generated spectrograms from the training utterances served as training samples for the custom Voice-CNN. During testing using the Spectrogram-Voting concept, the 5-second utterance length to be tested is split into 12 sub-utterances each of 0.4 seconds duration (the 13th sub-utterance is ignored because it is not up to 0.4 seconds). Each sub-utterance is converted to a spectrogram, resized to 128×128 pixels, and converted to grayscale. These spectrograms are fed to the trained custom Voice-CNN model which predicts whom each of the spectrograms belongs to along with the scores associated with the predictions. Since these spectrograms came from a single utterance, they must belong to a specific subject. The predicted candidates are then put to vote and the candidate most predicted (voted for) is deemed to be the owner of the utterance. Figure 4 illustrates this concept; the algorithm generated 12 spectrograms and counted the speaker labels predicted by the custom Voice-CNN for each of the spectrograms from a voice sample. 9 out of the 12 spectrograms indicated the voice sample belongs to candidate 1; therefore, the voice sample belongs to candidate 1 with the majority vote. There may be cases where the highest vote is equally divided between two or more subjects. The scores of all the votes for each affected candidate as predicted by the custom Voice-CNN model are computed and averaged to address this tie. The candidate with the highest average score is then assigned to be the owner of the utterance. The pseudocode for this concept is captured in Algorithm 1. The Spectrogram-Voting algorithm is evaluated based on the total number of correct utterances predicted relative to the total number of utterances tested.

FIGURE 4.

Illustration of spectrogram-voting for speaker identification.

Show All

Algorithm 1 Spectrogram-Voting Method for Speaker Prediction

Input:
Utterances from speakers

Output:
Speaker predictions for the utterances

```
For i = 1 to numberOfSpeakers do
    Generate: $S_{i1}$ , $S_{i2}$ , $S_{i3}$ ,...$S_{iN}$ %Spectrograms for speaker i
    Generate: $D_{i1}$ , $D_{i2}$ , $D_{i3}$ ,...$D_{iN}$ %Spectrogram decisions for speaker i
    Generate: $U_{i1}$ , $U_{i2}$ , $U_{i3}$ ,...$U_{iP}$ %Unique votes for speaker i
    Generate: $C_{i1}$ , $C_{i2}$ , $C_{i3}$ ,...$C_{iP}$ %Vote count for speaker i
    j = index at max($C_{i1}$ , $C_{i2}$ , $C_{i3}$ ,...$C_{iP}$)
    if length(j) > 1 %There is a tie in candidates with maximum vote
        For k = 1 to length(j) do
            $E_k$ = score($U_{ij(k)}$)
        End
        M = index at max(E)
        Prediction (i) = $U_{ij(M)}$
    Else
        Prediction (i) = $U_{ij(1)}$
    End
End For
```

Key: % indicates comments

2) Speaker Verification Using Spectrogram-Based Vote-Codes

Given any two utterances, voice verification is the process of determining if they belong to the same person or otherwise. This is usually done by generating voice prints or features for the utterances and using a distance metric to determine how similar the utterances are. In other words, verification is the process of determining if a speaker is who the person claims to be. The voice features used in this research are referred to as Vote-Codes; a novel concept based on the hypothesis that if a model has been trained for speaker identification using voice utterances belonging to subjects in a certain database, the training can be applied to subjects on a completely different database based on similarity ratios. Although the trained model might not have seen the user in the new database before, that new user will always be closer to the same set of people on which the model was trained. If the new user was predicted to be closer to x people in the training database, then all instances of testing the same new user will always correspond to the same x people in the training database. This set of x users is thus used in this research to generate an x-dimensional code which is referred to in this work as a Vote-Code (since these are the x indices of candidates most voted for by the trained network). Following the philosophy behind this hypothesis, a different person will have his or her own unique Vote-Code different from that of another subject.

A test spectrogram is resized to 128×128 pixels and converted to grayscale using image processing techniques. The trained model is used to generate a Vote-Code for the spectrogram. The Vote-Code was a 10-dimensional vector containing the indices of the 10 candidates in the training phase predicted to be the closest in similarity to the test spectrogram. In cases where the test utterance is long enough to generate more than one spectrogram, a Vote-Code is generated for each spectrogram. These codes are then combined into a single 10-dimensional Vote-Code. The elements in the individual codes that appear most are selected to fill the first entry of the combined code. This process is repeated for the next element in the individual vote codes that appears most until the elements in the combined Vote-Code are complete. In a situation where there is more than one element to be selected having a tie in the number of votes, the average of their scores as predicted by the trained network is computed and the element with the highest value is selected. This concept of generating Spectrogram-based Vote-Codes is captured in the flowchart of Figure 5.

FIGURE 5.
Flowchart for creating a vote-code.

Show All

Matching two Vote-Codes to determine if they belong to the same person boils down to a simple arithmetic set intersection rule, the more the set members the two vote-codes have in common, the more similar they are and vice versa. This hypothesis makes use of very short feature lengths (10) compared with x-vectors, i-vectors, or d-vectors (usually of length 128 or more) mainly used in literature and achieves state-of-the-art results despite not using machine learning or deep learning typical in literature for matching feature vectors. Specifically, (1) is used to discriminate between the utterances. If the number of elements in the intersection is greater than or equal to Ω (an adjustable threshold) then the utterances belong to the same person, otherwise, they belong to different subjects.

$$X \cap Y \geq \Omega \tag{1}$$

View Source where X and Y are Vote-Codes computed from two voice segments and Ω is the threshold for discrimination.

SECTION VII.

Fusion Design

Bimodal biometric designs make use of two traits for recognition such as the face and voice modality employed in this research. It has the potential to improve upon the best performance in unimodal cases. This section explores bimodal recognition systems by performing experiments with feature-level, score-level, and decision-level fusion.

A. Feature-Level Fusion

The feature-level fusion was a horizontal concatenation of the face and spectrogram image after conversion to grayscale as shown in Figure 6. Each of the 300 face and spectrogram images was combined this way and then split in the ratio 80:20 for training and testing of a custom Face-Voice-CNN. The structure of the Face-Voice CNN is the same as in Table 5, it uses an epoch of 8 and a learning rate of 0.002. The input layer was modified to accommodate the size of the new feature which was 128×256 pixels.

FIGURE 6.
Fusion of grayscale face and spectrogram images.

Show All

B. Score-Level Fusion

This research explores the combination of the scores of prediction from the custom Face-CNN and custom Voice-CNN. CNNs make their final predictions based on score ranking. The score for each candidate is the probability of the candidate being the unknown candidate in focus. The scores for all the candidates sum up to 1 (unity), and the candidate with the highest score probability is eventually predicted by CNN to be the unknown candidate.

1) Custom CNN Score-Level Fusion

Given a set of candidates to predict, a CNN generates a matrix of scores with the rows corresponding to the number of subjects in the group and the column equal to the number of candidates used during training. The candidate with the highest score in a row is the person predicted to be the unknown candidate queried for that row. In this case, the experiment combined the matrix of scores generated by the conventional use of both Face and Voice-CNNs by applying certain weights to control the contribution from each CNN to the fusion. Let the matrix of scores for the voice trait be V, described by (2) and the matrix of scores for the face trait be F, defined by (3).

$$V=\begin{bmatrix} P_{V11} & P_{V21} & \cdots & P_{VM1}\\ P_{V12} & P_{V22} & \cdots & P_{VM2}\\ \vdots & \vdots & \ddots & \vdots\\ P_{V1N} & P_{V2N} & \ldots & P_{VMN}\end{bmatrix} \tag{2}$$

$$F=\begin{bmatrix} P_{F11} & P_{F21} & \cdots & P_{FM1}\\ P_{F12} & P_{F22} & \cdots & P_{FM2}\\ \vdots & \vdots & \ddots & \vdots\\ P_{F1N} & P_{F2N} & \ldots & P_{FMN}\end{bmatrix} \tag{3}$$

View Source where P_{Vij} is the probability that the i_{th} individual queried is the j_{th} speaker, and P_{Fij} is the probability that the i_{th} face of the queried individual belongs to candidate j. The task is to solve the problem of allocating weights W_v (voice modality weight) and W_F (face modality weight) subject to the constraint described in (4), which maximizes the overall prediction accuracy, P, in (5).

$$W_V+W_F P=1=rowMaxIndex(W_V\times V+W_F\times F)(4)(5)$$

View Source

This work optimized the allocation of weights to the face and voice modalities through an iterative process captured in Algorithm 2.

Algorithm 2 Optimal Weight Determination for Score-Level Fusion

```
Input:
**Weight combinations**

Output:
**Prediction performance**

i = 0
For $W_v$ = 0 to 1 step 0.01
    a. i = i + 1
    b. Prediction (i) = rowMaxIndex($W_V \times V + W_F \times F$ )
End For
i = 0
For $W_v$ = 0 to 1 step 0.01
    a. i = i + 1
    b. matchCount(i) = match(Target, Prediction (i))
End For
```

```
$j$ = index at max(matchCount)
$W_v = j \times 0.01$
$W_F = 1 - W_v$
```

Key: V (Voice score), F (Face score), W_V (Voice weight), W_F (Face weight)

2) Spectrogram-Voting Score-Level Fusion

In this experiment, the matrix of voting scores obtained from the proposed Spectrogram-Voting algorithm replaces the scores from the custom Voice-CNN. The Spectrogram-Voting score was computed as the fraction of votes received by each candidate during the voting process. For a set of candidates to be predicted using the Spectrogram-Voting algorithm, the voting score matrix has a row equal to the number of people to be anticipated and a column length equal to the number of candidates used to train the custom Voice-CNN. The index of a particular individual queried corresponds to the row in the matrix while the elements represent the fraction of votes received by the candidates in the columns. The candidate with the highest fraction of votes is the predicted speaker. The voting score of the proposed Spectrogram-Voting algorithm for the voice modality was combined with the one from the custom Face-CNN by an optimal allocation of weights determined by the same concept as the case for custom CNN score-level fusion.

C. Decision-Level Fusion

The decisions of the spectrograms from an utterance are counted with that from the face. Figure 7 illustrates the concept of decision-level fusion. The 12 spectrograms together with the corresponding face of the subject give 13 images to be predicted. The prediction for each of the 12 spectrograms was done using the custom Voice-CNN model while the prediction for the face was carried out using the custom Face-CNN model. The votes from both modalities were then combined to make the final decision. However, for fairness, since 12 contributions came from the voice while only one came from the face modality, a voice modality weight, W_v and a face modality weight W_F were applied before the counting of votes subject to (7). Let the candidates voted for by the generated N spectrograms from a voice utterance be V_1, V_2, V_3,\ldots, V_N and the vote from the face sample be F_1. The study counts votes as described in (6) maximizing weight allocation using Algorithm 3.

$$P=\mathrm{mode}(W_V\times(V_1,V_2,V_3,\ldots,V_N),W_F\times F_1)\tag{6}$$

View Source .

FIGURE 7.
Illustration of decision-level fusion.

Show All

Algorithm 3 Optimal Weight Determination for Decision-Level Fusion

Input:
Weight combinations

Output:
Prediction performance

```
h = 0
For W_v = 0 to K step 1
    h = h + 1
    For I = 1 to totalSpeaker
        Prediction (I,h) = $\mathrm{mode}(W_V\times (Vi_1, Vi_2, Vi_3,\ldots, Vi_N ), (K - W_V)\times Fi_1)$
    End For
End For
h = 0
```

```
For $W_v$ = 0 to K step 1
    h = h + 1
    For I = 1 to totalSpeakers
        matchCount(I,h) = match(Target, Prediction (I,h))
    End For
End For
$W_v$ = h at max(matchCount)
$W_F$ = K – $W_v$
```

Key: V (Voice score), F (Face score), W_V (Voice weight), W_F (Face weight)

The statistical mode was employed in (6). The weights are subject to the constraint of (7).

$$W_V+W_F=K \quad (7)$$

View Source where W_v, W_F, and K are integers.

The study chose the optimal weights through an iterative process of observing performance as weights were varied using Algorithm 3.

SECTION VIII.

Results and Discussion

The statistics of NaijaFaceVoice were earlier captured in Tables 1–4. Its characteristics are depicted in Figure 8, showing its variability in blur, rotation, occlusion, scale, translation, illumination, and pose. The voice counterpart was also characterized by channel and background noise as well as channel mismatches owing to the different devices from which the videos were uploaded to YouTube. This makes the created database suitable for research capturing real-life challenging scenarios. The face and voice counterparts of NaijaFaceVoice were evaluated in terms of relative purity with the Tufts face database [49] and VoxCeleb, respectively. The model generated when NaijaFace was trained on the custom Face-CNN and the one for Tufts face on the same CNN were used in turn to perform face verification on the Extended Yale face database [50] independent of NaijaFace and Tufts face. The same approach was used to evaluate NaijaVoice relative to VoxCeleb on the Nigerian mini database [6]. This approach to evaluation is not new, it has been applied by [14] and [51]. This research, however, puts a value to it in terms of relative purity using (8) and (9), as shown at the bottom of the page,

Face_DB_Relative_PurityVoice_DB_Relative_Purity=NaijaFace_DB_AccuracyRef_Face_DB_AccuracyRef_Face_DB_Accuracy×100%=NaijaVoice_DB_AccuracyRef_Voice_DB_AccuracyRef_Voice_DB_Accuracy×100%(8)(9)

View Source for evaluating NaijaFace and NaijaVoice respectively. The evaluation was done for matching thresholds (Ω) set to 1 and 2 using the Vote-Code concept and the results averaged. At other thresholds, there was no match for both databases being compared. The evaluation result is summarized in Tables 6 and 7. It is seen from these tables that NaijaFace and NaijaVoice were estimated to be relatively purer by 0.7% and 0.68% respectively. These results show that NaijaFaceVoice can reliably be employed for related research works.

FIGURE 8.
Sample faces and voice segments in the database depicting variability.

Show All

A. Performance of Custom Face-CNN Recognition System

The performance of the designed unimodal recognition system using the custom Face-CNN is shown in Table 8 with the EER plot in Figure 9. The difference in features is compared to a threshold to match how close two faces are. If the difference is below the threshold, both faces are deemed to belong to the same person, otherwise, they are different. The higher this threshold, the more likely two traits being compared are deemed to be similar and the more the FAR. The lower this threshold, the stricter the recognition system and the more the FRR. FAR and FRR are thus relative to the set threshold, and can vary between two possible extreme values [52]. However, the parameter more universal is the EER which is the point of intersection of these two curves at 0.16%.

FIGURE 9.
EER plot for face recognition based on custom Face-CNN.

Show All

B. Performance of Voice Recognition Systems

In Table 9, the custom CNN-based method for speaker recognition attained an accuracy of 79.67%. The proposed Spectrogram-Voting method increased speaker identification accuracy to 96.86%, an improvement of 17.19%. The EER also shows the superiority of the proposed Spectrogram-Voting method over the custom CNN-based method by reducing EER from 11.20% to 1.36%, reducing error by a factor of 8. The EER plot for the custom CNN-based and the Spectrogram-Voting method for speaker verification are shown in Figure 10 and Figure 11, respectively.

FIGURE 10.
EER plot for speaker recognition based on custom Voice-CNN.

Show All

FIGURE 11.
EER plot for speaker recognition based on Spectrogram-Voting.

Show All

The power of the concept of Spectrogram-Voting is illustrated using the spectrograms of frames from the utterance of an individual. In this work, there were 12 such spectrograms put to vote. Suppose 9 of these spectrograms predicted candidate 1 as the owner of the utterance while the other 3 predicted other different candidates. Assuming candidate 1 was the correct individual and the identification was made at the frame level, then the accuracy would be 9/12 or 75% (since 3 spectrograms were predicted wrongly out of 12). However, in the Spectrogram-Voting concept, because the majority of the spectrograms predicted candidate 1, the algorithm is surer that at the utterance level, the voice belongs to candidate 1 and assigns the utterance to that candidate. This utterance was correctly predicted (1/1) giving an accuracy of 100% which is an improvement of 25% over the custom use of CNN.

C. Performance of Feature-Level Fusion

Table 10 and Figure 12 show the performance evaluation of testing the design based on feature-level fusion in terms of performance metrics and EER plot, respectively. It achieved an accuracy of 99.94%, an improvement of 0.27% over the best result achieved in the unimodal cases with the face recognition system. In addition, the EER of 0.028772% is an improvement of over five times the best result achieved in the unimodal instances with the best EER from the face recognition system.

FIGURE 12.
EER plot for feature-level fusion.

Show All

D. Performance of Score-Level and Decision-Level Fusion

Table 11 captures the evaluation results for custom CNN score-level fusion, Spectrogram-Voting-based score-level fusion, and Spectrogram-Voting-based decision-level fusion. The performance of these three methods surpassed the best result of the unimodal cases using both TSR and EER. The research realized these optimized results by the best appropriation of weights, which controlled the extent to which the face and the voice modality contributed to the fusion. All three experiments used an iterative process described in Algorithms 2 and 3 to obtain the best allocation of weights. The results of the iterations in the case of custom CNN score-level fusion, Spectrogram-Voting-based score-level fusion, and Spectrogram-Voting-based decision-level fusion are shown in Figure 13, Figure 14, and Figure 15, respectively. In all three plots, the best weight was the point on the weight axis at which the curve peaked. The corresponding plots for the FAR and FRR, with their point of intersection depicting the EER for custom CNN score-level fusion, Spectrogram-Voting-based score-level fusion, and Spectrogram-Voting-based decision-level fusion, are shown in Figure 16, Figure 17, and Figure 18, respectively.

FIGURE 13.
Plot of performance vs weights for custom score-level fusion.

Show All

FIGURE 14.
Plot of performance vs weights for spectrogram-voting score-level fusion.

Show All

FIGURE 15.
Plot of recognition performance vs weights for decision-level fusion.

Show All

FIGURE 16.
EER plot for custom CNN score-level fusion.

Show All

FIGURE 17.
EER plot for spectrogram-voting score-level fusion.

Show All

FIGURE 18.

EER plot for decision-level fusion.

Show All

It is seen that the decision-level fusion based on the proposed Spectrogram-Voting method achieved the best result. This proposed method also shows the effectiveness of the algorithm. It attained an identification accuracy of 99.98%, an improvement of 0.31% over the best unimodal design achieved with the face modality alone. More interestingly, it achieved an EER of 3.519×10^{-4} %, unprecedented in the literature on NaijaFaceVoice. This improvement is 450 times better in terms of error reduction than the best result in the unimodal cases. The second best result was score-level fusion which was also based on the proposed Spectrogram-Voting concept achieving an EER of 3.318×10^{-3} .
The reason for these improvements is attributed to a deeper interaction between the voice spectrograms and the face image, the process of making all the spectrograms within utterances contribute not only indicates the potential frames that belong to the right individual but also eliminates the misleading spectrograms thereby reduces errors. To further showcase the superiority of the proposed Spectrogram-Voting concept for speaker identification, and the novel Vote-Code concept for speaker verification, they were benchmarked with other related state-of-the-art methods on the popular VoxCeleb database. As seen in Table 12 and Table 13, the proposed method in this work outperformed the state-of-the-art for speaker verification, and speaker identification, respectively.

Table 14 answers the question of the extent of improvement in performance because of the fusion of the face and voice traits by comparing the better of the two results in the unimodal cases with the result of fusion. As seen from the results, improvements in literature are marginal, however, the robust fusion method used in this work significantly improved performance by reducing EER by over a factor of 450. This improvement is attributed to the availability of a large-scale database (NaijaFaceVoice) allowing CNNs to learn the variability in the traits and thus better cope with the variability in these traits. The proposed Spectrogram-Voting method shown to outperform state-of-the-art is also another reason because the spectrograms within an utterance can better cooperate. Again, these spectrograms were made to associate closely with the face traits. The use of iterative methods to optimize weight allocation to the fusion contributed to the outstanding fusion performance.

SECTION IX.

Conclusion

The focus of this research was to create a large-scale face-voice database to be made publicly available owing to the scarcity of such databases and to significantly improve recognition performance through a robust fusion of both traits. The research adopted a semi-automated approach to mitigate against errors in NaijaFaceVoice because some state-of-the-art databases have been discovered to contain errors. The semi-automated pipeline was used to curate the faces and voices of Nigerians on YouTube to populate NaijaFaceVoice. NaijaFaceVoice contains over 2 million faces and more than 140,000 utterances up to 195 hours for 2656 Nigerian subjects labeled for biological gender and language. NaijaFaceVoice was estimated to be as clean as state-of-the-art databases that are error-free. This work introduced the concept of Spectrogram-Voting to improve speaker recognition which outperformed state-of-the-art on VoxCeleb using the shortest feature length so far recorded in literature. The Spectrogram-Voting concept was extended to generate very compact fixed-length Vote-Codes irrespective of utterance length. Vote-Codes are highly discriminative voice features for speaker verification. With the aid of a custom Face-CNN and custom Voice-CNN, the proposed Spectrogram-Voting concept was incorporated into a score-level and decision-level fusion of face and voice attaining a record EER of 0.0003519% on NaijaFaceVoice.

ACKNOWLEDGMENT

This research is fully sponsored by Covenant University Centre for Research, Innovation and Development (CUCRID), Covenant University, Ota, Nigeria.

Authors
Figures
References
Keywords
Metrics

Speech Recognition of Sundanese Dialect Using Convolutional Neural Network Method with Mel-Spectrogram Feature Extraction

2023 11th International Conference on Cyber and IT Service Management (CITSM)

Published: 2023

Show More