Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca

Ogechukwu Iloanusi, Ugogbola Ejiogu, Ife-ebube Okoye, Ijeoma Ezika, Samuel Ezichi, Charles Osuagwu
Department of Electronic Engineering
University of Nigeria, Nsukka 410001, Nigeria
e-mail: ogechukwu.iloanusi@unn.edu.ng, ugogbola@gmail.com, ifeebube.okoye@unn.edu.ng, ijeoma.ezika@unn.edu.ng, ezichisam@gmail.com, charles.osuagwu@unn.edu.ng

Emenike Ejiogu
Department of Electrical Engineering
University of Nigeria
Nsukka 410001, Nigeria
e-mail: emenike.ejiogu@unn.edu.ng

***Abstract*—Voice verification and gender classification from voice were carried out in the context of native (mother tongue) languages and lingua franca languages. A total of 3980 voice utterances recorded in English language and 28 native languages were acquired from 520 bilingual subjects in this paper. We first determined the cross linguistic influence of mother tongue by bilingual speakers on the verification performance of voice recognition using English and native languages' gallery and probe sets. Secondly, we employed transfer learning in training four convolutional neural network models for classifying gender from voice, using training and test samples of English language, exclusively; one dominant native language; and a mixture of 28 native languages. Our results do show that mother tongue or first language, intonation variations, language variety in the training or test sets do influence voice verification and gender classification.**

Keywords-voice; verification; gender classification; accuracy; mother tongue; native language; lingua-franca; intonation

I. INTRODUCTION

Voice recognition has been in existence in the literature and implemented in systems for biometric person recognition; access control; mobile communications and home automation, to name a few applications. In biometric-based person recognition, which involves - verification or classification, a robust voice recognition system should be implementable in several environments and across languages without its quality being compromised. The performance of voice biometric systems could be influenced by several factors such as noise [1], speech duration [2], language and intonation variations [1][3]. This makes the selection of feature vectors difficult and sensitive.

The effect of language variations in voice biometrics has been studied over the years. A comparison of two voice recognition systems was carried out in [4] suggesting that voice recognition systems were language dependent. The studies carried out in [3]–[5] state that language affects voice recognition but failed to highlight the impact of language variations on voice biometrics. Intra-class intonation variations negatively impact performance in voice recognition systems [1], [3], [6]. Intonation refers to a combination of acoustic parameters such as pitch and intensity accompanying human speech, which could vary according to the emotion and intention of the speaker. It has also been researched that varying speech texts and lengths in a person results in linguistic complexities and varying phonological properties in speeches [7].

With regard to the factors that impact voice recognition, namely, language and intonation, we analyze the effect of the speakers' first language or mother tongue and second language on the performance of voice recognition and gender classification from voice. It is known that a mother tongue is spoken fluently, competently, intuitively and with natural emotions and intonations [8], [9]. On the other hand, a second language often lacks the attributes of a mother tongue language – intuition, competence, fluency [10]. A speaker's mother tongue often influences the intonation of the spoken second language [11]. Also, a language used as a lingua franca, becomes very dynamic, and assimilates new accents, forms and linguistic characteristics as it spreads across other native languages [12].

There are as many languages as there are tribes in Nigeria, because each tribe has its native language, which each people of the tribe naturally speak as their mother tongue and first language. The link language or lingua franca for Nigerians is English language, which most educated Nigerians speak with varying confidence levels. However, most Nigerians are bilingual and speak their mother tongue fluently, and the mother tongue intonation largely affects the spoken English.

We first determine the influence of mother tongue versus lingua-franca languages on voice recognition by carrying out voice verification experiments. The error rates resulting from voice verification experiments carried out in this paper are meant to show how the test sets are influenced by lingua-franca and mother tongue languages. Secondly, via machine learning experiments, we evaluate how training sets composed of training samples of (1) English language set of utterances (2) a native language set of utterances (3) and a mixture of native language set of utterances influence the classification of gender from voice.

Gender classification when integrated in biometric identification expedites recognition. It narrows the search space to a specific gender, and thus expedites data retrieval through biometric indexing. It also provides useful clues for

978-1-7281-4577-8/19/$31.00 ©2019 IEEE

crime scene investigation. It can be applied in E-commerce in online shopping of accessories that are gender specific.

In this paper, we determine how mother tongue, lingua franca, intonation, language variety, affect voice verification and gender classification from voice using training and test sets of utterances made using the speakers' first language - mother tongue and second language – English. We:

1. Compare performances of verification of a subject's gallery to a probe in the context of lingua franca and Native languages. We determine the cross-linguistic influence of mother tongue versus lingua-franca by bilingual speakers on performance of voice recognition.
2. Compare performances of classification of a subject's gender from voice prints of lingua franca – English language; one dominant native language; and 28 native languages. In other words, we determine how the accuracy of gender classification from voice is influenced by training and test sets of voice utterances made up of second languages and native languages.

This paper is organized as follows: The paper is introduced and relevant works reviewed in Section I. The dataset acquired in this paper is described in Section II. Feature extraction methods used in this paper for voice verification experiments and gender classification are explained in Section III. Experiments and results are discussed in Section IV.

II. Data Preparation

A total of 3980 voice utterances were acquired from 520 Nigerian subjects in two sessions between June and November 2018. The subjects comprise 326 male and 194 female subjects with 1964 and 2016 voice utterances, respectively. Five different speeches of English and the equivalent translated native languages were acquired from the subjects in the first and second sessions. More females participated in the second session hence their samples were more in number. Considering that the participants in the data collection, Nigerians, are bilingual, all speak English and at least one other native language, preferably, their mother tongue. The number of native languages collected in the voice utterance dataset is 28, namely, Afemai, Akoko-Edo, Annang, Efik, Ekoi, Fulani, Hausa, Ibibio, Idoma, Igala, Igbo, Igede, Ijaw, Ika, Ikom, Ikwerre, Ishan, Kaire-Kaire, Kanuri, Lokaa, Urhobo, Yoruba, Obudu, Ogoni, Okobo, Okirika, Tiv and Ukwani. The dominant tribe amongst these native languages in the dataset is Igbo.

The voice set comprises two major sets - the English and 28 native language sets. Each of the two sets has a gallery (Session 1) and probe (Session 2) set. The English language set has utterances of five different speeches in the gallery and probe sets. The native language gallery and probe sets have utterances of the translated equivalent of the five English language texts. The text of the English utterances and the average timing per utterance per gallery and probe sets are shown in TABLE 1. The 28 native translations of each of the 5 English texts would be too bulky if provided in the paper and readers would not appreciate them. The timing for each utterance is computed by dividing the number of audio samples in an utterance by its sampling frequency.

TABLE I. English Texts and Average Timing (in Seconds) for Each of the Texts Calculated from the Sampling Frequency

Texts in each English speech set	*Gallery (secs)*	*Probe (secs)*
1. I am hungry and I am thirsty	1.97	1.93
2. I am from Nigeria	1.30	1.28
3. I am a student of the university	1.71	1.71
4. What else shall I say?	1.34	1.35
5. Please come and help me	1.32	1.33

III. Feature Extraction

Voice recognition techniques abound in biometric literatures. Hidden Markov Model (HMM), Neural Network Model (NNM), Dynamic Time Warping (DTW), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Relative Spectral Filtering (RASTA) and Vector Quantization (VQ), Genetic Algorithm (GA), Random Forest Recursive Feature Elimination (RF-RFE) have been used for voice recognition in the literatures [13]–[16]. Deep learning methods have been employed in [17] and by using Mel frequency cepstral coefficients (MFCC) in [18].

A. Extraction of Voice Features

In this paper, the Mel frequency discrete wavelet cepstral coefficients were used (MFDWC). Discrete wavelet transform (DWT) based on Debauchees wavelet was used on the signal energies of the coefficients computed through a bank of 40 filters tuned to the Mel scale.

An utterance is divided into two or 224 batches, B, before extracting features for the voice verification or gender classification tasks, respectively. The Fast Fourier transform (FFT) of each B of N samples, $X(k)$, and consequently, the power spectrum, P_k, of the $X(k)$ are obtained. A filter-bank of 40 frequencies in the Mel scale, centred between 20 Hertz and the Nyquist frequency, F_N, is computed. $F_N = {}^{F_s}/_{2}$, where F_s is the sampling frequency of the digitized utterance. The Mel of a frequency, $M(F)$ is $1125\left(\ln\left(1+\frac{F}{700}\right)\right)$. $M(20)$ and $M(F_N)$, are determined, consequently, the values of 40 equally spaced filters between $M(20)$ and $M(F_N)$ are computed. P_k is mapped to the 40 equally spaced filter banks in the Mel scale, M, with centre frequencies between $M(0)$, and $M(F_N)$. The logarithm of the weighted sum of the resulting coefficients is computed, resulting in a matrix of $B \times 40$. That is, 2×40 in the case of voice recognition and 224×40 in the case of gender classification from voice. Computing the DWT of the $B \times 40$ matrix, using the Debauchees wavelet, results in a $B \times n$ matrix, were n depends on the levels of wavelets used.

B. Convolutional Neural Network Architecture

Traditional and machine learning approaches using convolutional neural network (CNN) were combined for gender estimation from voice in this paper. However, CNN method is appropriate for images. Transfer learning was used to train four CNN models for the gender classification tasks

using the VGG16 pretrained model [19]. The last three layers of the VGG16 were replaced with a fully connected layer of size 2, softmax and two-way classification layer for gender estimation. The VGG16 model is shown in Figure 1. This model has an input layer size of $224 \times 224 \times 3$, five convolutional blocks and fully connected layers. We transformed MFDWC features to appear as images to the machine learning algorithm and CNN model. Each voice utterance was divided into four bundles in order to have as many training samples for machine learning. In keeping with the input size requirement, each bundle was divided into 224 batches and MFDWC features were extracted.

A total of 224 coefficients were extracted from the computation of the Debauches discrete wavelet transform, up to the tenth level, on the each of the 224 batches mapped to 40 Mel filters. The output of each bundle was a matrix of size 224×224 with values from the set of $\mathbb{R}$ numbers. Given that values in each matrix are to 4 decimal places and could be negative, each value in the samples, s_v, was normalized to a scale of positive integer values between 0 and 255, thus, representing the matrix as a grayscale image as follows:

$$s_v \longleftarrow M\{s_v + round(s_{v_{MAX}} - s_{v_{MIN}}) + C\} \quad (1)$$

where C is a constant and M is a scaling factor dependent on the range of the MFDWC values. Each matrix was augmented to a 3-channel matrix of size, $224 \times 224 \times 3$.

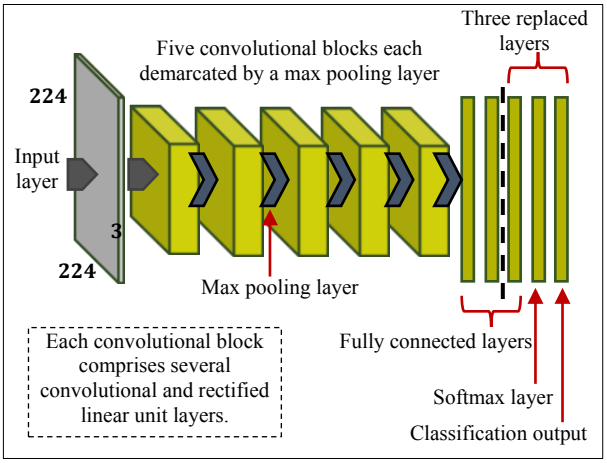

Figure 1. Architecture of CNN for the gender from voice models.

IV. Experiments, Results and Discussions

A. Categories of Experiments in the Paper

The gallery and probe sets were defined from the audio utterances acquired in the first and second sessions. Two (2) experimental tasks were designed for voice verification and classification experiments as follows:

I. Five probe and Five gallery sets were defined where each set has English and native languages. The first task comprises twenty experiments.

II. Gender was estimated from voice notes of samples of (1) English utterances made by all subjects (2) Native language utterances made by all subjects (3) English language utterances made by the dominant native tribe (4) Native language utterances made by the dominant native tribe. There are four experiments in the second task.

Experiments in this paper were carried out in MATLAB, python and a Macintosh operating system with a 16 GB RAM. GPU resources were used for training models.

B. Biometric Evaluation and Classification Performance Metrics

Biometric evaluations were carried out in this paper using the commonly used performance metrics, namely, false accept rate (FAR), false reject rate (FRR), genuine accept rate (GAR), equal error rate (EER) and receiver operating characteristics (ROC) [20]. An exhaustive comparison of all subjects' features was carried out yielding match scores – genuine and impostor. FAR, FRR, GAR and EER were computed based on varying decision thresholds, T, between the minimum and maximum match scores as follows:

$$FAR|_T = \frac{number\ of\ false\ accepts\ |_T}{total\ number\ of\ impostors} \quad (2)$$

$$GAR|_T = \frac{number\ of\ genuine\ accepts\ |_T}{total\ number\ of\ genuines} \quad (3)$$

Varying thresholds, rather than a single threshold, removes the bias of fixing an error rate on a single preferred threshold in biometric evaluations. EER is the error rate value at which FAR equals FRR. The less the EER the better the algorithm. ROC is a plot of GAR versus FAR for all varying thresholds. EER and ROC were eventually derived for each evaluation based on the set of FAR, FRR and GAR.

The first experimental task has twenty EERs and ROCs. However, results are presented as area under the ROC curve (AUC) in order to minimize text and graphs in this paper. AUC is a value that varies between worst performance (0) and best performance (1). Results of the second experimental task of gender classification are presented using precision rates $\boldsymbol{P}$, recall rates $\boldsymbol{R}$, and overall accuracy $\boldsymbol{A}$, expressed as:

$$P = \frac{Number\ of\ correct\ predictions\ for\ a\ gender}{Number\ of\ all\ predictions\ for\ a\ gender} \times 100 \quad (4)$$

$$R = \frac{Number\ of\ Correct\ predictions\ for\ a\ gender}{Ground\ truth\ number\ for\ a\ gender} \times 100 \quad (5)$$

$$A = \frac{Summation\ of\ all\ correct\ predictions}{Total\ number} \times 100 \quad (6)$$

C. Task I: Experiments and Results – Comparison of Voice Verification Performances of various Speech Texts of English and Native Languages Utterances

The experiments in this task comprise four English and four native gallery sets, four English and four native probe

sets. An English utterance set in this experiment was compared with other four English utterance sets, likewise a native utterance set is compared with other four native utterance sets. Following the method of experiment described in Section IV (A), there are 20 evaluations resulting in 20 AUCs and EERs shown in TABLE II.

The results in TABLE II have varying AUC / EER results for the different utterances' comparisons. The results also appear to vary consistently for both English and corresponding translated native language utterances. The results - AUCs / EERs, in the native utterances comparisons (rows 5 to 8) are ordered according to their corresponding English utterances comparisons (rows 1 to 4) in TABLE II.

Nine of the results in rows 5 to 8, for the native utterances' comparisons, outperform the corresponding English utterances' comparisons in rows 1 to 4 in TABLE II. The AUCs of the evaluations involving native utterances in rows 5 to 8 are higher, and EERs lower, than the corresponding English utterances (rows 1 to 4), except in the comparisons between native speeches 1 and 5. The higher AUCs and lower EERs results in rows 5 to 8 could be due to the competence and fluency of the subjects in speaking their mother tongue compared to speaking English language.

TABLE II. VOICE VERIFICATION PERFORMANCES FOR ENGLISH AND NATIVE LANGUAGES UTTERANCES PRESENTED AS AUC AND EER

No		*Probe sets*			
	Gallery sets	*English 2*	*English 3*	*English 4*	*English 5*
		AUC (EER)	*AUC (EER)*	*AUC (EER)*	*AUC (EER)*
1.	**English 1**	0.6553 (37.02%)	0.6468 (39.82%)	0.6053 (41.10%)	0.6110 (40.60%)
2.	**English 2**		0.6710 (38.03%)	0.6352 (39.60%)	0.6195 (41.45%)
3.	**English 3**			0.6847 (35.07%)	0.6516 (37.82%)
4.	**English 4**				0.7053 (34.86%)
		Native 2	*Native 3*	*Native 4*	*Native 5*
5.	**Native 1**	0.6876 (35.95%)	0.6569 (37.80%)	0.6433 (38.27%)	0.6093 (41.77%)
6.	**Native 2**		0.7122 (35.39%)	0.6869 (35.97%)	0.6655 (38.11%)
7.	**Native 3**			0.7281 (33.61%)	0.6791 (36.91%)
8.	**Native 4**				0.7231 (33.68%)

There is a consistent pattern in variation in results amongst English and native utterances' comparisons in TABLE II. The comparisons of the utterances between sets 2 and 3, 3 and 4, 4 and 5, in TABLE II, produced the best results in both the English and native language sets'. The worst results are those between sets 1 and 4, 1 and 5, in both English and native utterances' comparison. This consistent pattern in results could be due to the varying average speech length in the utterance sets as stated in TABLE 1.

In summary, the results show that the performance of voice recognition is improved in the context of mother tongue or first languages, due to the competence and fluency of the subjects in speaking their mother tongue. Hence, speeches involving more intonation and linguistic rhythm, as in the mother tongue spoken naturally, may improve voice recognition. Secondly, variation in speech length of gallery and probe sets could affect voice recognition.

D. Task II: Experiments and Results – Gender Classification from Voice Models Trained with English and Native Utterances

Each recorded speech was divided into four voice utterances in order to have as many training and test samples as possible. The subjects and samples in the training set and test sets are disjointed. Hence, testing of models is entirely on new subjects and samples.

Four models, m_1, m_2, m_3 and m_4, listed in TABLE III were trained in this experimental task using training and test samples set of (i) English voice utterances made by subjects from all 28 natives (ii) utterances comprising 28 languages made by all natives (iii) English utterances made by one dominant native tribe (iv) utterances of one native language made by one native tribe, respectively.

The sets in (iii) and (iv) comprises utterances made by a group of only one dominant tribe - Igbo. The Igbo tribe comprises 458 persons out of a total of 520 persons that participated in the data collection. In the dominant tribe there are 278 males and 180 females with 3112 and 3080 training samples, respectively.

The same training parameters were used in all models trained in this paper. Minibatch size, learning rate, momentum and L2 regularization were set to 256, 0.0001, 0.9, and 0.0001, respectively. The training parameters for the four trained models, m_1 through m_4, are provided in TABLE III. The models were validated with disjointed subjects' samples in the training. Validation accuracies are shown in TABLE III. The accuracy of the gender estimations for the four test samples were computed and shown in TABLE IV.

TABLE III. ATTRIBUTES OF TRAINING SAMPLE SIZE, TEST SIZE AND TIME FOR FOUR GENDER CLASSIFICATION FROM VOICE MODELS

M	*Train / Test set*	*Train / Test sample size*	*Test size (male / female)*	*Training time*	*Validation accuracy*
m_1	(i)	7,120 / 792	380 /412	02:14:19	83.3%
m_2	(ii)	7,112 / 792	380 /412	02:10:55	84.8%
m_3	(iii)	6,192 / 792	396 /396	01:53:41	82.3%
m_4	(iv)	6,192 / 792	396 /396	01:56:45	85.1%

The ground-truth test set sizes for the male and female gender are provided in TABLE III. The results in TABLE IV show a precision of 85.84%, 86.52%, 85.16% and 85.21% for the male gender and 81.39%, 83.49%, 79.91% and 85.75% for the female gender, for all evaluations of (***i***) on $\boldsymbol{m_1}$, (***ii***) on $\boldsymbol{m_2}$, (***iii***) on $\boldsymbol{m_3}$ and (***iv***) on $\boldsymbol{m_4}$, respectively. Models, m_1, m_2 and m_3, generate more false positives for the female gender. The evaluations yield recall rates of 78.16%, 81.05%. 78.28% and 85.86% for the male gender and 88.11%, 88.35%, 86.36% and 85.10% for the female gender. The recall rates are higher for the female gender in m_1, m_2 and m_3, showing that these three models estimate the female gender better than the male.

The overall accuracy of (***i***) on $\boldsymbol{m_1}$ is 83.33% in TABLE IV while m_2 yields an overall accuracy of 84.85%. Sets (***i***) and (***ii***) have the same training samples' size, test samples' size and same subjects, but differ in the language set – English language for ***(i)*** and 28 native languages for ***(ii)***. Overall accuracy in m_2 is better than the accuracy obtained from model m_1. The slight increase in performance shows that in voice gender classification, training and test sets of utterances made in the native language would improve accuracy.

Overall accuracy in m_3 and m_4 are 82.32% and 85.1%, respectively, in TABLE IV, hence, m_4 outperforms m_3. m_4 was trained and tested with a set of native voice utterances made by one dominant tribe – set (***iii***), while m_3 was trained and tested with English language utterances – set (***iv***). (***iii***) and (***iv***) have the same training samples' size, test samples' size and same subjects, but differ in language only. m_4 has the best accuracy amongst the four trained models.

The results in TABLE IV show that accuracy is improved when a native group speaks their mother tongue or first language, supporting the results and deductions drawn from earlier experiments in Section IV (C). In summary, performances of gender classification from voice are improved with utterances of a mother tongue language compared to a second language.

TABLE IV. TRAINING SAMPLE SIZE, TEST SIZE AND TIME FOR FOUR GENDER CLASSIFICATION FROM VOICE MODELS

Results	***(i) on*** $\boldsymbol{m_1}$	***(ii) on*** $\boldsymbol{m_2}$	***(iii) on*** $\boldsymbol{m_3}$	***(iv) on*** $\boldsymbol{m_4}$
Number of predicted males	346	356	364	399
Number of predicted females	446	436	428	393
Number of males correctly predicted	297	308	310	340
Number of females correctly predicted	363	364	342	337
Precision male	85.84	86.52	85.16	85.21
Precision female	81.39	83.49	79.91	85.75
Recall male	78.16	81.05	78.28	85.86
Recall female	88.11	88.35	86.36	85.10
Overall accuracy	83.33	84.85	82.32	85.48

V. CONCLUSION

We determined and evaluated the influence of mother tongue and intonation variations, on probe and gallery sets of utterances acquired in English and 28 native languages from the same subjects. We also evaluated the impact of one language, cross-linguistic effects, mother-tongue and mixed language variety on gender classification from voice on four convolutional neural network (CNN) models trained using transfer learning in this paper. Our results reveal several lessons. First, performance of voice recognition is improved in the context of mother tongue, due to the competence and fluency of the subjects in speaking their mother tongue. This shows that speech involving more linguistic rhythm and intonation in one language may improve recognition. Secondly, variation in speech length of gallery and probe sets could affect voice recognition. Finally, result of gender classification from voice is optimal when training sets have a mother tongue language rather than a second language.

REFERENCES

[1] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.

[2] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3244–3252, Jul. 2018.

[3] M. FarrÚs and Mireia, "Voice Disguise in Automatic Speaker Recognition," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–22, Jul. 2018.

[4] R. Auckenthaler, M. J. Carey, and J. S. D. Mason, "Language Dependency in Text-Independent Speaker Verification."

[5] G. Boulianne, "Language-independent voice passphrase verification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4490–4494.

[6] K. Sebastian and L. Mary, "FASR: Effect of voice disguise," in *2016 International Conference on Emerging Technological Trends (ICETT)*, 2016, pp. 1–4.

[7] A. Martin and S. Peperkamp, "Assessing the distinctiveness of phonological features in word recognition: Prelexical and lexical influences," *J. Phon.*, vol. 62, pp. 1–11, May 2017.

[8] N. Calet, N. Gutiérrez-Palma, and S. Defior, "A cross-sectional study of fluency and reading comprehension in Spanish primary school children," *J. Res. Read.*, vol. 38, no. 3, pp. 272–285, Aug. 2015.

[9] X.-Y. Gao, "A Comparison of First and Second Language Acquisition," in *Humanity and Social Science*, 2017, pp. 281–288.

[10] W. Q. YOW, J. S. H. TAN, and S. FLYNN, "Code-switching as a marker of linguistic competence in bilingual children," *Biling. Lang. Cogn.*, vol. 21, no. 5, pp. 1075–1090, Nov. 2018.

[11] S.-A. Jun and M. Oh, "Acquisition of Second Language Intonation," in *Sixth International Conference on Spoken Language Processing* , 2000, pp. 73–76.

[12] A. Mauranen, "English as a global Lingua Franca: changing language in changing global academia," *Explor. ELF Japanese Acad. Bus. Context.*, vol. 33, no. January 2015, pp. 29–46, 2015.

[13] Nisha, "Voice Recognition Technique : A Review," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 5, no. V, pp. 262–268, 2017.

[14] H. S. I. Harba and E. S. I. Harba, "Voice Recognition with Genetic Algorithms," *Int. J. Mod. Trends Eng. Res.*, vol. 2, no. 12, pp. 144–155, 2018.

[15] K. Zvarevashe and O. O. Olugbara, "Gender Voice Recognition Using Random Forest Recursive Feature Elimintion with Gradient Boosting Machines," in *2018 International Conference on Advance in Big Data, Computing and Communication systems (icABCD)*, 2018, pp. 1–6.

[16] Ranny, "Voice Recognition using KNN and Double Distance Method," in 2016 International Conference on Industrial Engineering, Management Science and Application (ICIMSA), 2016, pp. 1–5.

[17] D. Polap and M. Wozniak, "Image Approach to voice recognition," in *IEEE Symposium Series on Computational Intelligence(SSCI)*, 2017, pp. 3207–3213.

[18] H.-S. Bae, H.-J. Lee, and S.-G. Lee, "Voice Recognition Based on Adaptive MFCC and deep learning," in *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, 2016, pp. 1542–1546.

[19] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[20] T. Dunstone and N. Yager, *Biometric System and Data Analysis: Design, Evaluation, and Data Mining,* . New York, USA: Springer Science LLC, 2009.