Person Identification System from Speech and Laughter Using Machine Learning Algorithms

Oluwatoyin P. Popoola, Comfort O. Folorunso, Olumuyiwa S. Asaolu, John J. Joshua and M. D. Oyeyemi
Department of Systems Engineering, University of Lagos, Akoka, Lagos, Nigeria
E-mail: opopoola@unilag.edu.ng

Abstract
Automated person identification and authentication is paramount for preclusion of cybercrime, national security and veracity of electoral processes. This is a critical component of Information and Communication Technology (ICT), which is the mainstay for national development. This paper presents the use of speech and laughter of people for person identification with the focus on forensics application where people speak and laugh in between. Features were extracted using the Librosa library in Python programming language via Scientific Python Development Environment (SPYDER) IDE (version 4.1.3) of the Anaconda software. While the Orange software (version 3.25.0) for data-mining was used for training, testing and validation of five standard machine learning algorithms: Neural Networks (NN), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB) and Logistic Regression (LR). Results showed that the neural networks classifier gave the best accuracy followed by the SVM. There was an average of 17.6% and 14.1% increase in the validation metrics when both speech and laughter were combined as compared to speech and laughter independently respectively. This research area is very useful in forensics especially for recognising criminals in conversation.

Keywords: *person identification, laughter, speech, forensics*

1.0 INTRODUCTION

Human voice is a vital occurrence which is highly reliant on its producer. No two people have exactly the same voices as every voice has a specific frequency range which defines a person's voice. The vocal cord can be in several positions depending on what is being said at a particular time. It can also be affected by the position of the articulators which include the tongue, teeth, lips and palate amongst others. Since the arrangement, size, shape and movements of these articulators are different for every individual, then there is no way the voice of two individuals can be exactly the same (Kinnunen and Li., 2009; Mokgonyane *et al*., 2019). Differences amongst speakers may include variations in the vocal cords and its shape as well as variations in speaker's expression and speaking styles (Mokgonyane *et al*., 2019). Different speakers have their unique pattern and choice subset for laughing, the duration of laughter units (i.e. phones and airflow phases) can also be characterised with individuals (Urbain and Dutoit, 2011). According to the study by Devillers and Vidrascu (2007), the articulate expression of laughter changes across sex, individuals and settings.

Speaker identification also known as speaker recognition is a fast-growing research direction in speech signal processing. It is concerned with the problem of identifying a person from his/her voice characteristics. This problem evolves in so many applications such as personal authentication in E-commerce systems, recognizing persons in a conversation for forensics, and security check in military environments. Speaker recognition system performs two major roles, these are Speaker Identification (SID) and Speaker Verification (SV). In SID system for instance, the question: "Who said what?" is being answered while a true/false question is answered in SV system, such as "Did a particular person gave a specific comment or not?" (El-Ayadi *et al*., 2017).

Speaker Verification (SV), which is also known as voice/person verification/authentication, is a way of determining if the speaker identity is who the person claims to be. It is a one-to-one identification system. The major elements of a verification system are front-end processing; speaker modelling and pattern matching. In contrast, SID is a way of finding the identity of an unknown speaker by matching his/her voice features with that of registered speakers stored in the database. It is a one-to-many identification system. The major element of SID is the same as the SV, the main difference being that the speaker models are saved in parallel form while the most-likely person is reported (Feng, 2004).

Both speaker identification and verification system are divided into text-dependent and text-independent system. In text-dependent speaker recognition systems also known as closed set system, speakers are permitted to say particular utterances, phrases or words which must be the same with those saved in the database and the unknown voice must also come from a set of known speakers that are saved in the database. While in text-independent recognition systems also known as open set system, the speakers are free to say any utterance, speech or word of their choice and the unknown voice may come from unregistered speakers (Feng, 2004; Bakmand-Mikalski *et al*., 2007). Many researchers have carried out studies on speaker recognition using speech signal (Mokgonyane *et al*., 2019; Sun *et al*., 2019; Yasmin *et al*., 2019; Singh and Joshi, 2020). Also, research in laughter analysis has gone from automatic laughter detection in meeting (Laskowski and Schultz, 2008) to laughter synthesis (Cakmak *et al*., 2014) and recently to person identification (Folorunso *et al*., 2020).

Laughter is one of the most imperative paralinguistic events, and it has an important role in human conversation. The automatic detection of laughter incidences in human speech can support automatic speaker recognition systems and paralinguistic task such as speaker emotion detection as well as identifying humorous content in video clips. Integrating laughter detection in automatic speech recognition (ASR) systems can assist in minimizing word error rate by recognizing non-speech sounds (Gosztolya *et al*., 2016).

Speech has extensively been used over the years to recognise an individual while laughter is just being introduced for this purpose. There are some differences between speech and laughter. Laughter is highly variable as compared to speech; because the glottal configuration of laughter is different from speech due to high subglottal pressure. Also, speech is controlled while laughter can be voluntary (acted) as well as involuntary (spontaneous). Hence, it cannot be controlled once it is aroused. Laughter, therefore, may be more beneficial when established for forensic use. Speech/voice can be mimicked perfectly but laughter can hardly be mimicked. When both laughter and speech are employed for person identification, there is tendency of high accuracy of recognising who said what in a forensic scenario, such that if the suspect mimicked someone's voice, he/she can still be recognised from his/her laughter. For instance, Ruch *et al*., (2019) in a recent study, agreed that there are biometric traits in laughter like in fingerprints and confirmed that this field has not been exploited. Hence, the aim and major contribution of this study is therefore, to investigate the use of both speech and laughter signature for person identification in conversation for forensic application *especially for recognising criminals in conversation*.

2.0 RELATED WORK

With numerous developments in technology, the security sector has encountered important advancement. We can say that technology has transformed security but the challenge was how to make computers identify or recognise an individual. Biometric being based on measuring physical and behavioural characteristics uses features, which are commonly and readily available to all classes of people. These features are distinct, easily collected and tested for, and have high variability to carter for repetition of data class. Some of the biometric features currently in use include: Facial thermogram, hand vein, gait, keystroke, odour, ear, hand geometry, fingerprint, face, retina, iris, palmprint, voice, signature and DNA. Delac and Grgic as well as Folorunso *et al*., presented a concise summary of different biometric methods which include single and multiple biometric systems. Voice-based biometric system uses some of the features of human-speech that are invariant for a particular individual. Though the behavioural features of the same human speech varies over time due to age, medical, emotional as well as environmental conditions. The voice-based biometric system is classified into automatic speaker verification and identifications. The automatic speaker verification system uses voice as validation characteristic in a one to one verification scenario. While the automatic speaker identification system uses voice to recognise who a person truly is. A particular voice feature of an individual is matched against a stored pattern in a database. A typical voice feature can be formants or any other sound characteristics which are unique to each individual's vocal tract (Delac and Grgic 2004; Folorunso *et al*., 2019).

Tran *et al*., (2004) introduces normalization technique which depends on fuzzy set theory to improve the performance of voice-based verification. In order to authenticate a claimed personality, a likeliness value was evaluated with a threshold in order to allow or reject the person. The use of noise clustering as well as the c-means clustering membership function was introduced to eradicate the problem of ratio-type scores, which affects the false acceptance rate. Their findings showed great reduction in the false acceptance and false rejection rate. Also, in order to classify information from an audio system, Sheikh *et al*., (2020) generated a fuzzy muting function utilizing the first and second set of variables and the corresponding score related to each sub-word. Their method was effective for the purpose of audio signal information muting system. Working from a neuro-cognitive point of view, Belin *et al*., (2004) looked at neural association of voice perception. The ability to evaluate a person's gender and age bracket from listening to their voices was a very strong motivation behind their work. They tried to look at the voice as an auditory face in comparison with the face recognition system. They then suggested the use of Bruce and Young's model of face perception as a structure for understanding the perceptual as well as the cognitive processed contained in voice perception. This suggested model predicts functional detachment similar to those perceived for faces. In addition, Nagels *et al*., (2020) researched into children's discrimination and weighting for voice gender categorization. Their findings showed that children's capabilities to differentiate and consider voice cues for categorisation takes some times to develop. Subsequently, Krawczyk and Jain (2005) considered the large evolution from paper-based medical records towards electronic medical records, guaranteeing the security of such private and highly sensitive data cannot be over-emphasized since the health care practitioners only need to edit and update patient's record on the tablet or even smart phones, hence the need to protect the patient's privacy as required by all governmental regulations. A safe authentication system must be put in place anytime such records are to be accessed, and as such the need for biometric-based access cannot be disputed. Research showed that online signature integrated with voice modalities is the most appropriate means for the users in such a verification system since tablet is built with the associated devices. More so, Shakil *et al*., (2020) introduced a biometric authentication and data control system for healthcare information in cloud. Finding revealed that this method was a very effective storage and retrieval electronic-healthcare system as the speedup recorded was nine times the existing methods.

Voice is a more natural means of communication. Verbal communication is quicker and more efficient than textual communication. Bhogal *et al*., (2012), evaluated the use of virtual universe (VU) residents also known as Avatars in online service employing audio biometrics. They used voiceprint to approve operation limited to an authorise user. In summary, they demonstrated the possibility of using biometric in internet-based activities. Scheffer *et al*., (2013) on the other hand tried to deal with two of the challenges facing voice biometrics technology. These problems consist of non-ideal recording conditions which are often operational situation challenges such as noise, echoes, voice channels and audio compression. In order to improve on the existing system, Jagdale *et al*., (2020) introduced a robust speaker recognition system by combining low-level spectral features and prosodic features. Their findings recorded 15 – 20% improvement accuracy.

Subsequently, Kaur and Kaur (2016), presented a brief evaluation of different voice biometric for speaker verification in attendance system. They proposed the use of voice, having considered different methods that have been employed for automatic attendance for student and as such the use of voice for this purpose is a very important and highly welcome phenomenon. They used gammatone filter bank instead of Mel filter bank, after which the discrete cosine transform was applied to separate overlaying signals. The use of Gammatone frequency cepstral coefficient (GFCC) with the Gaussian Mixture Model and Artificial Neural Networks (ANN) was incorporated for training and matching task respectively.

Considering the technological improvement of the social media, whereby some other people use these social media as a form of terrorism to transmit their message. A typical biometrics recognition method such as face or fingerprints has been substituted by another biometric trait such as voice as this may be readily available in such scenario. Mazaira-Fernandez *et al*., (2015) introduced a gender-dependent extended biometry factors (GDEB). The GDEB factors classify features extracted from voice source, tract factors and other pertinent features such as format data, having in mind that male and female voices show both acoustic-phonetic variations as well as physiological differences. The main idea was to improve classification rate in speaker recognition using few parameters. Lowe *et al*., (2020) presented a structured review spanning over a decade of studies using speech for automated assessment of psychiatric disorder. Their findings showed that speech processing technology could aid mental health assessment though there might be some challenges to overcome.

Furthermore, Vatsa *et al*., (2009) having researched into various biometric technology, identified some serious problems affecting this technology and categorised them into accuracy, computational speed, security, cost, real-time attacks and scalability. They also identified the various possible attacks on biometric technology and these include impersonation, coercive, replay attack, as well as the attack on feature extractor, template database, matcher and matching results amongst others. In improving the performance of biometric technology, they identified two major ways one can protect the biometric information from such attack. These include encryption as well as watermarking. Nagakrishnan and Revathi (2020) presented a multiple chaotic maps and Deoxyribonucleic acid (DNA) encryption technique for a robust speech encryption-based person authentication. Their results showed that the encryption system resists the brute force, differential and statistical attack. Farzaneh and Toroghi (2020) on the other hand presented a novel watermarking technique using graph-based transform (GBT) and singular value decomposition (SVD). Their findings showed that the suggested method has a high resistance to different attacks. Korshunov and Marcel (2016) considered the fact that most biometric technology systems are vulnerable to spoofing which reduces their wide use, hence they presented the need to develop anti-spoofing detection methods also refer to as presentation attack detection (PAD) systems. They presented an integration of PAD and Automatic Speaker Verification (ASV) systems. Also, Kamble *et al.*, (2020) presented a comprehensive literature review of the various spoofing challenges encountered by the automatic speaker verification. These challenges include synthetic speech, voice conversion, replay, twins and impersonation. Chettri et al (2020) on the other hand researched the impact of different sub-bands with their importance on replay spoofing detection using two benchmark spoofing datasets: ASVspoof 2017 dataset and ASVspoof 2019 PA datasets. The presented sub-band Convolution Neural Network (CNN) model performs better than the traditional full-band CNN model.

In order to improve on speaker verification/identification systems, some researchers such as Arora and Vig (2020), Tawara *et al*., (2020) and Rohdin *et al*., (2020) used short utterances rather than the usual long utterances employed in literature which consist of silence periods. Their results were more accurate compared to the existing methods. Also, some researchers tried to improve on this task by improving on existing features commonly used. For instance, Jahangir et al. (2020) and Abd El-Moneim *et al*., (2020) combined time-based features and spectrum or log-spectrum with the traditional Mel-Frequency Cepstral Coefficients (MFCC) respectively, some improvement in the classification accuracy were recorded. However, no study has been reported where both speech and laughter of individuals are combined for person identification system. Hence the major contribution of this work.

3.0 METHODOLOGY

From various literature studied, there was no suitable dataset for this task, hence the need for building our own dataset. Data were obtained from a local event dataset built for the purpose of biometric analysis from the University of Lagos Laughter dataset. This data consists of both speech and laughter of 70 individuals (56 males and 14 females) as they read some phrases and laughed, each with minimum of 15 audio samples of speech and laughter combined as well as 20 speech and laughter samples independently. The signal was passed through a high-pass filter removing lower frequencies signals, in order to mimic the human voice where the vocal tract nearly operate like a high pass filter (Bakmand-Mikalski *et al*., 2007). The data was denoised in the PRAAT® software using the spectral subtraction method (Boersma and Weenink, 2015). Different features were extracted using the Librosa library in Python programming language via the Scientific Python Development Environment (SPYDER) IDE (version 4.1.3) of the Anaconda software (Mc-Fee et al., 2015). The dataset is available at http://laughter-db.herokuapp.com/.

Figure (1) shows the flow diagram of the methodology used in this study.

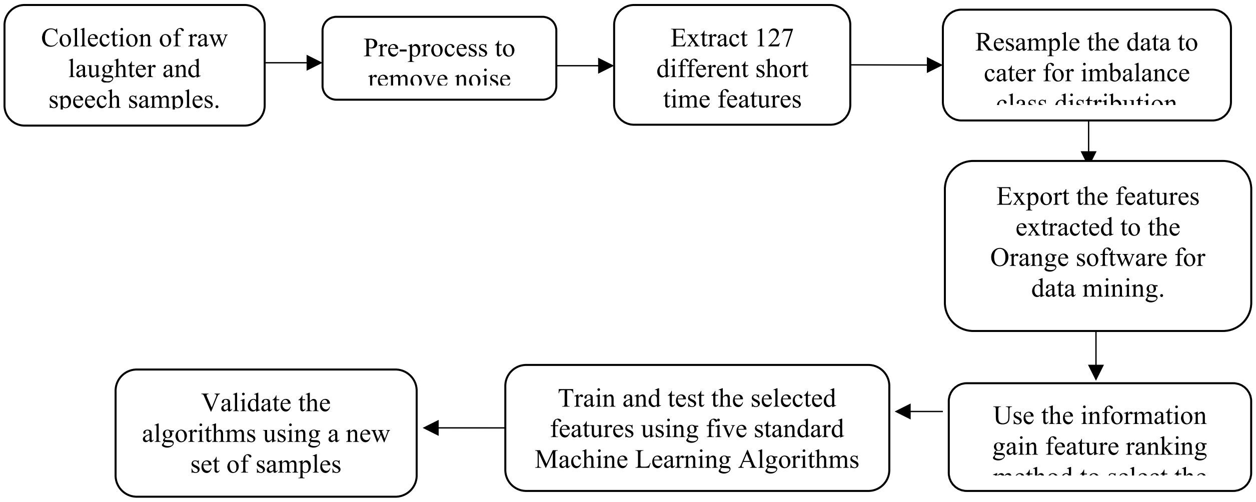
Collection of raw laughter and speech samples.
Pre-process to remove noise
Extract 127 different short time features
Resample the data to cater for imbalance class distribution
Export the features extracted to the Orange software for data mining.
Use the information gain feature ranking method to select the
Train and test the selected features using five standard Machine Learning Algorithms
Validate the algorithms using a new set of samples

Figure 1: Flow diagram of the methodology

The features extracted include statistical features, acoustic features, time-domain features, prosodic features, frequency domain features and cepstral-domain features. A total of 127 different features were extracted. These features were described in Tables 1 to 6. Due to variation in speaking length and laughter duration of each individual, the dataset used in this study is imbalanced in terms of the number of samples in each class. Such imbalance of class distribution may cause the model to classify towards the majority class, hence the need to resample the data in order to generate a balanced class for effective fitting of the machine learning algorithms on the transformed datasets. The SMOTETomek library in the imbalanced-learn Python library was used to achieve this. Over-sampling method is used to generate a set of synthetic samples in the minority class while under-sampling method remove or combine samples in the majority class (Jason, 2020).

These features were exported to the Orange data mining environment (version 3.25.0) in the anaconda software (Adekitan and Salau, 2020). The information gain feature ranking method was then used to select the best features which were used to train five standard machine learning algorithms. These include Support Vector Machine (SVM), Neural Networks (NN), Random Forest (RF), Logistic Regression (LR) and Naïve Bayes (NB). The models were trained and validated on a 10-fold cross-validation. The workflow for the training and validation in the Orange software is shown in Figure (2). To evaluate the proposed system, speech and laughter samples were used independently from the same dataset for the analysis, extracting the same number of features for training, testing and validating with the five standard machine learning algorithms in the Orange data mining environment.

Table 1: Statistical features

Feature ID	Feature Name	Description	Definition	References
1	Signal mean (μ)	The statistical mean evaluated from the audio signal over the period of the low-level framework.	$\mu = \sum_{n-1}^{N} X[n]$	Thoman, (2009)
2	Signal Std (δ)	The statistical standard deviation evaluated from the audio signal over the period of the low-level framework.	$\delta = \sqrt{\sum_{n=1}^{N} (X[n] - \mu)^2}$	Thoman, (2009)
3	Signal skewness (Y_1)	Skewness is the balance distribution of the probability density function (PDF) of the amplitude of a time series.	$Y_1 = \frac{\sum_{n=1}^{N} (X[n] - \mu)^3}{\delta^3 (N-1)}$	Thoman, (2009)
4	Signal kurtosis (Y_2)	Kurtosis is a statistical measure that quantifies the distribution shape of a signal with respect to a Gaussian distribution.	$Y_2 = \frac{\sum_{n=1}^{N} (X[n] - \mu)^4}{\delta^4 (N-1)} - 3$	Thoman, (2009)

Where, *N* is the number of samples in the low-level analysis frame; X[n] is the sample value at sample index n.

Table 2: Acoustic features

Feature ID	Feature Name	Description	Definition	References
5 – 24	Formant frequencies F1-F5 (mean, min, max, range)	These are the resonant frequencies of the vocal tract.	$F_1 = \frac{c}{4L}$	ResearchGate, (2014)

Where c is speed of sound, L is the acoustic length.

Table 3: Time-Domain Features

Feature ID	Feature Name	Description	Definition	References
25	Tempo	Tempo is the rate of pulse represented by the inverse of the beat period.	$Tempo = \frac{1}{Beat\ Rate}$	Scheirer, (1998)
26-27	Root Mean Square (RMS) - Mean and Std	RMS is the square root of the mean square of the signal	$\mathrm{RMS} = \sqrt{\frac{1}{N} \sum_{n-1}^{N} (X[n])^2}$	Thoman, (2009)
28-29	Zero crossing rate (ZCR) – mean and Std	It is the rate of sign – changes between the values of two successive samples in an audio signal.	$\mathrm{ZCR}_t = \frac{1}{N} \sum_{n=1}^{N} \lvert sign(x_t[n]) - sign(x_t[n-1]) \rvert$	Thoman, (2009)

Where, x_t is the sample value at index of the audio signal at frame index t.

Table 4: Prosodic features

Feature ID	Feature Name	Description	Definition	References
30 -34	Pitch (mean, max, min, range and std)	Also called fundamental frequency, it is the lowest or principal frequency in a periodic signal.	$F_0 = \frac{1}{T}$	Bäckström, (2019)
35 - 39	Intensity (mean, max, min, range and std)	This is the measure of the energy or loudness of a signal. It is related by the square of the amplitude.	$I = \frac{(\Delta p)^2}{2\rho v_w}$	Urone and Hinrichs, (2020)

Where Δp is the change in pressure amplitude (N/m^2), ρ is the density of the material in which the sound wave travels (Kg/m^3), v$_w$ is the speed of sound in the medium (m/s).

Table 5: Frequency-domain features

Feature ID	Feature Name	Description	Definition	References
40 - 41	Spectral-centroid (mean, std)	This is the center of gravity of the magnitude of the frequency domain spectrum representation of an audio signal.	$SC_t = \frac{\sum_{n=1}^{N} M_t[n]F[n]}{\sum_{n=1}^{N} M_t[n]}$	Thoman, (2009)
42 - 47	Spectral-bandwidth – 2,3,4 (mean, std)	This shows if the frequency band energies are concentrated around the spectrum centroid or dispersed across the entire spectrum.	m_n $= \int (x - \mu)^n . p(x) dx$	Thoman, (2009)
48-61	7 Spectral-contrast (mean, std)	This is the difference between peaks and valleys in the spectrum.	$Peak_k$ $= \log (\frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x'_k , i)$ $Valley_k$ $= \log (\frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x'_k N - 1$ $+ 1)$ $SC_k = Peak_k - Valley_k$	Jiang, (2002)
62-63	Spectral-roll-off (mean, std)	This is the frequency below which a certain stipulated percentage of the overall spectrum magnitude distribution is focused.	$\sum_{n=1}^{N(SR_t)} M_t[n] = R \sum_{n=1}^{N} M_t[n]$	Thoman, (2009)
64-87	12 Chroma-stft (mean, std)	This is 12-element description of the spectral energy.	$\{C, C^{\#}, D, D^{\#}, E, F, F^{\#}, G, G^{\#}$	Kattel et al., (2019)

Where, N is the number of frequency bands in the spectrum representation of the signal, F[n] is the frequency represented by band index n, $M_t[n]$ is the magnitude of the spectrum at frame index t and band index n. $N(SR_t)$ is the band index of the roll-off frequency SR_t, R is the roll-off percentage.

Table 6: Cepstral-domain features

Feature ID	Feature Name	Description	Definition	References
88 - 127	Mel Frequency Cepstral Coefficient (MFCC) – mean and std	This is the cepstral description with frequency band distribution using the Mel-scale.	$c_n = \sum_{m=1}^{M} [\log Y\ (m)] \cos\ [\frac{\pi n}{M}(m - \frac{1}{2})]$	Mokgonyane, (2019)

Where n is the index of a cepstral coefficient, Y (m) is the output of an M-channel filter-bank for m = 1… M.

All the five machine learning algorithms used in this study are easy to implement and highly efficient when used to solve classification problems.

Support Vector Machine (SVM) is one of the popular machine learning classifiers. It is a supervised learning algorithm employed in classification and regression task and uses the concept of boundary to classify between classes (Jain et al., 2020). The polynomial kernel was implemented in this study.

Neural Networks (NN) is a computational machine learning model that is inspired by biological neural networks which is the central nervous system in human brain. It consists of input, hidden and output layers. It is also implemented in both classification and regression task (Palo et al., 2020).

Random Forest (RF) is a supervised machine learning algorithm used for classification and regression task. It uses multiple random decision trees, such that each tree is built on a random sample from the raw data and at each tree node, a subset of features is randomly selected to produce the best split (Fromont et al., 2020).

Logistic Regression (LR) is a Machine Learning algorithm which is used for classification tasks; it is a predictive analysis algorithm that uses the concept of probability for its analysis. It is used to predict a discontinuous outcome based on variables which may be discontinuous, continuous or mixed. Thus, when the dependent variable has two or more discontinuous outcomes, logistic regression is a commonly used technique. The outcome could be in various forms such as Yes / No, 1 / 0, True / False, High/Low, given a set of independent variables (Levitan et al., 2016).

Naïve Bayes (NB) is a simple, effective and popularly used machine learning algorithm. It a probabilistic classifier that implements the Maximum A Posteriori decision rule in a Bayesian network. It is used for classification task (Assuncao et al., 2020).

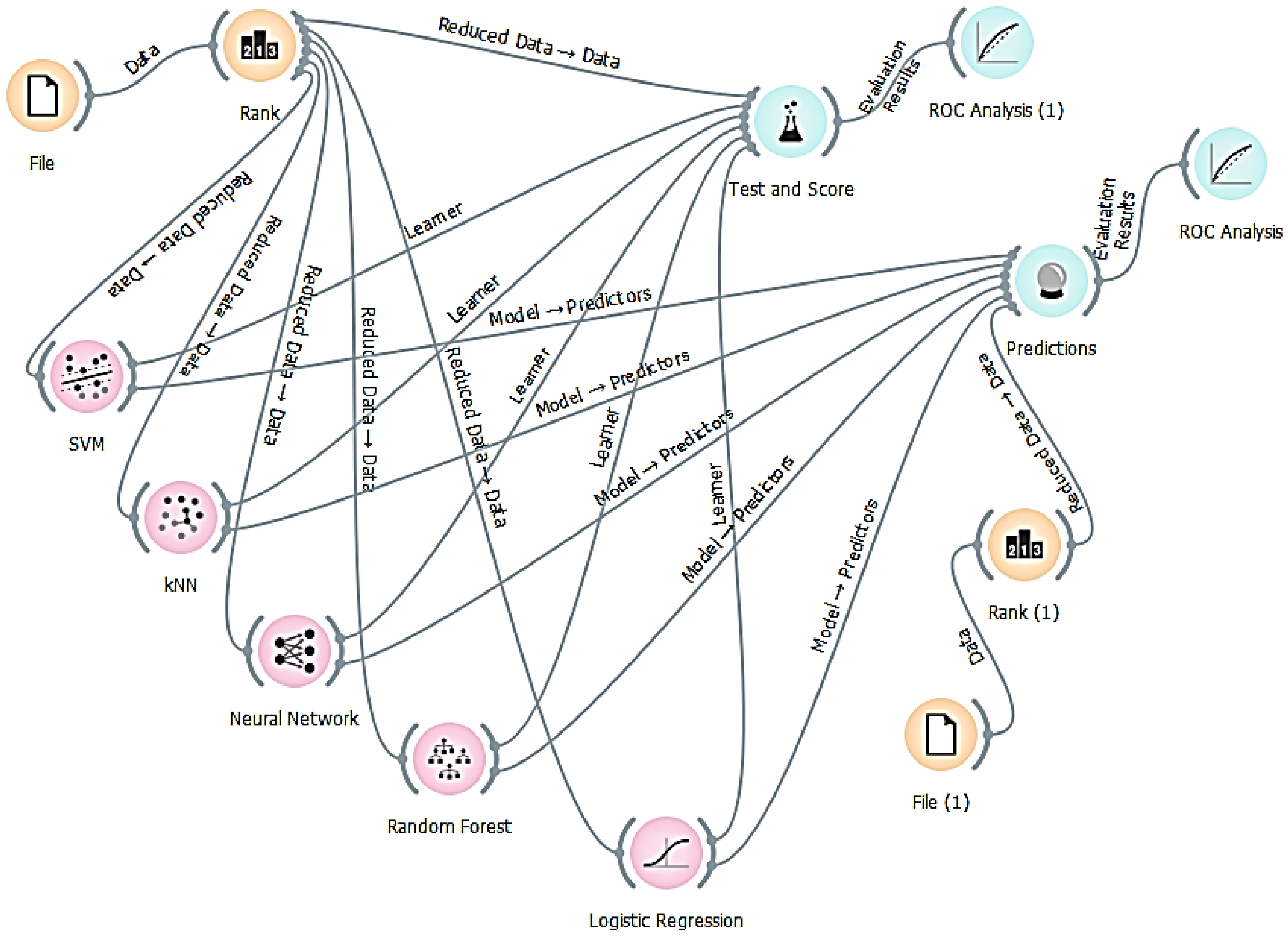

Figure 2: Workflow for training and validation of the machine learning algorithms in Orange environment

4.0 RESULTS AND DISCUSSION

The speech and laughter samples were both used for training the models. After numerous training, testing and validation via the Orange software while changing the feature ranking size, 80 best features gave the highest testing and validation accuracy. These best features are reported in Table 7. The results of the training and testing with the best features are shown in Table 8. New set of samples were then introduced to the system via the prediction widget and their validation result is shown in Table 9. The ROC curves for training and validation are shown in Figures (3) and (4). Speech and laughter were independently used to train the models, while new sets of samples were introduced to validate the models. Results of the training, testing and validation are shown in Tables 10 - 13.

Table 7: Eighty (80) best features from information gain ranking system used in training and testing the machine learning algorithms

Feature	Ranking value	Features	Ranking value	Features	Ranking Value
rmse_mean	1.563	mfccs_2_mean	1.147	F5_min	0.926
spectral_contrast_7_mean	1.556	zcr_mean	1.135	pitch_std	0.926
signal_mean	1.513	spectral_constrast_1_std	1.133	mfccs_18_std	0.925
signal_std	1.500	Pitch	1.128	chroma_stft_10_mean	0.925
Intensity	1.482	spectral_constrast_4_mean	1.122	mfccs_14_std	0.912
spectral_bandwidth_3_mean	1.458	spectral_constrast_6_std	1.092	spectral_contrast_3_mean	0.910
spectral_bandwidth_4_mean	1.458	mfccs_3_mean	1.088	spectral_rolloff_std	0.909
spectral_constrast_6_mean	1.444	mfccs_3_std	1.087	mfccs_17_std	0.907
spectral_bandwidth_3_std	1.442	mfccs_15_mean	1.067	mfccs_14_mean	0.906
spectral_bandwidth_4_std	1.442	mfccs_18_mean	1.064	mfccs_5_mean	0.894
rmse_std	1.442	mfccs_20_mean	1.044	F2_Hz	0.889
mfccs_1_mean	1.438	mfccs_8_mean	1.024	mfccs_11_std	0.888
spectral_bandwidth_2_mean	1.400	mfccs_8_std	1.022	F3_Hz	0.884
intensity_max	1.333	mfccs_9_mean	1.019	mfccs_13_mean	0.882
spectral_contrast_5_mean	1.332	F5_Hz	1.006	mfccs_20_std	0.881
intensity_range	1.328	mfccs_9_std	1.001	mfccs_15_std	0.860
intensity_min	1.319	mfccs_7_mean	0.989	mfccs_6_std	0.856
spectral_contrast_1_mean	1.307	F4_Hz	0.977	mfccs_16_mean	0.855
spectral_contrast_7_std	1.259	spectral_centroid_std	0.958	chroma_stft_2_mean	0.852
spectral_rolloff_mean	1.234	chroma_stft_12_mean	0.958	mfcc_4_std	0.852
mfccs_10_mean	1.224	mfccs_19_mean	0.948	mfccs_6_mean	0.849
spectral_centroid_mean	1.217	mfccs_7_std	0.941	spectral_contrast_2_std	0.848
intensity_std	1.208	F5_range	0.938	mfccs_13_std	0.846
mfccs_2_std	1.188	mfccs_10_std	0.938	mfccs_12_mean	0.835
spectral_bandwidth_2_std	1.179	chroma_stft_11_mean	0.932	mfccs_19_std	0.832
mfccs_17_mean	1.172	chroma_stft_1_mean	0.932	mfccs_11_mean	0.824
mfcc_1_std	1.161	spectral_contrast_5_std	0.931		

Table 8: Testing accuracy for both speech and laughter with the machine learning algorithms

Model	AUC	Classification Accuracy	F1_Score	Precision	Recall
SVM	1.000	0.991	0.991	0.991	0.991
RF	0.999	0.963	0.963	0.963	0.963
NN	1.000	0.994	0.994	0.994	0.994
LR	1.000	0.974	0.974	0.975	0.974
NB	1.000	0.977	0.977	0.978	0.977

Table 9: Validation accuracy for both speech and laughter on a new set of data

Model	AUC	Classification Accuracy	F1_Score	Precision	Recall
SVM	0.992	0.900	0.887	0.903	0.900
RF	0.947	0.686	0.664	0.726	0.686
NN	0.992	0.915	0.905	0.919	0.815
LR	0.982	0.628	0.614	0.726	0.628
NB	0.990	0.827	0.818	0.875	0.827

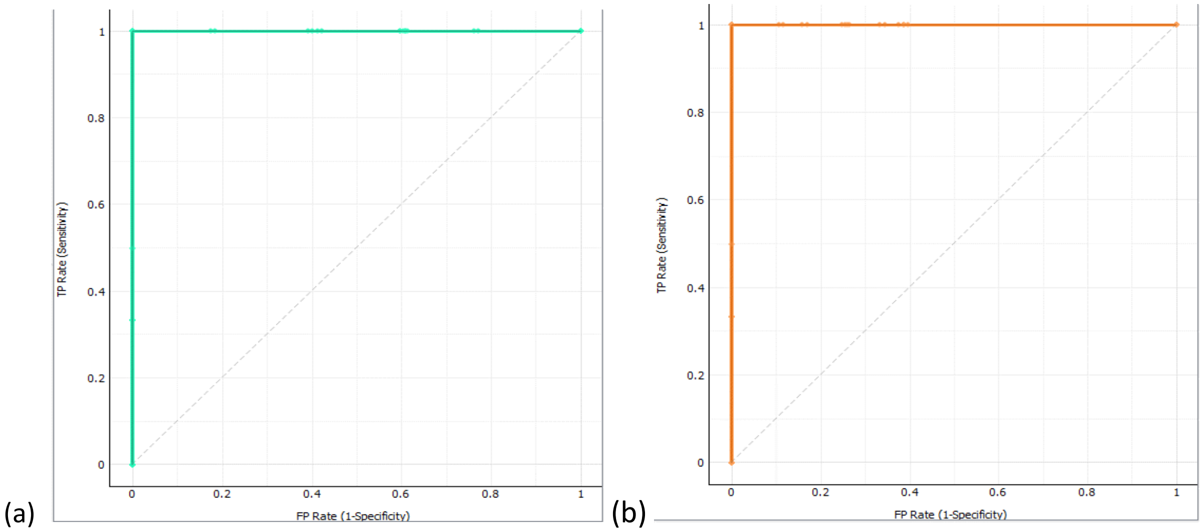

Figure 3: ROC curve for the training of (a) Support Vector Machine and (b) Neural Networks

All the ROC curves are the same for the training and testing of all the machine learning algorithms.

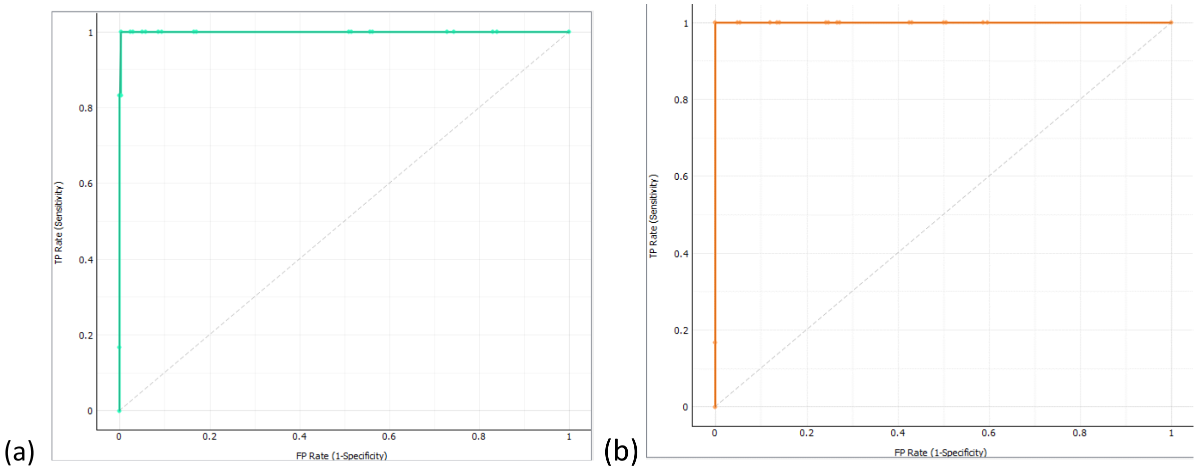

Figure 4: ROC curve for the validation of (a) Support Vector Machine and (b) Neural Networks

All the ROC curves are the same for the validation of all the machine learning algorithms.

Table 10: Testing accuracy for speech only with the machine learning algorithms

Model	AUC	Classification Accuracy	F1_Score	Precision	Recall
SVM	1.000	0.987	0.987	0.988	0.987
RF	0.997	0.951	0.951	0.952	0.951
NN	1.000	0.990	0.990	0.990	0.990
LR	0.999	0.965	0.964	0.966	0.965
NB	1.000	0.966	0.965	0.967	0.966

Table 11: Validation accuracy for speech only on a new set of speech samples

Model	AUC	Classification Accuracy	F1_Score	Precision	Recall
SVM	0.999	0.901	0.896	0.920	0.901
RF	0.950	0.571	0.542	0.613	0.517
NN	0.999	0.813	0.800	0.862	0.813
LR	0.995	0.726	0.711	0.780	0.756
NB	0.966	0.519	0.474	0.514	0.519

Table 12: Testing accuracy for laughter only with the machine learning algorithms

Model	AUC	Classification Accuracy	F1_Score	Precision	Recall
SVM	1.000	0.964	0.964	0.966	0.964
RF	0.995	0.921	0.920	0.922	0.921
NN	1.000	0.981	0.981	0.982	0.981
LR	0.998	0.929	0.929	0.931	0.929
NB	0.999	0.940	0.940	0.943	0.940

Table 13: Validation accuracy for laughter only on a new set of laughter samples

Model	AUC	Classification Accuracy	F1_Score	Precision	Recall
SVM	0.997	0.904	0.904	0.925	0.904
RF	0.969	0.745	0.744	0.802	0.745
NN	0.998	0.914	0.909	0.932	0.914
LR	0.980	0.615	0.577	0.624	0.615
NB	0.996	0.863	0.858	0.890	0.8633

Table 7 shows the 80 best features selected by the information gain ranking system. The performance report of the five standard machine learning algorithms used with both laughter and speech: area under ROC curve (AUC), classification accuracy, F1_score, precision and recall are shown in Table 8. Results showed that neural networks gave the highest AUC, classification accuracy, F1_score, precision and recall followed by the SVM, naïve Bayes, Logistic regression and lastly random forest. Similarly, the validation report is given in Table 9 and it also showed similar trend in the result with the neural network having highest performance metrics followed by SVM, naïve Bayes, random forest and lastly the Logistic regression. The neural network reported 92% classification accuracy, (0.91) F1_score, (0.92) precision and (0.82) recall. The high F1_score, precision and recall showed that the model did not overfit. The receiver operator characteristic (ROC) curve was closed to 1 for all the machine learning algorithms both for training and testing (Figure 3) as well as validation (Figure 4). The performance of our person identification system using both speech and laughter outperformed the results recorded by Gyanendra et al. (2011) and Medikonda *et al.,* (2020) with the standard Voxforge speech dataset. Gyanendra et al. (2011) reported 74.7% classification accuracy with sixty speakers while Medikonda *et al.,* (2020) reported an average of 79.26% classification accuracy.

From Table 10, there was a slight improvement (an average of 1%) in the training and testing performance metrics of all the classifiers when both laughter and speech was used as compared to speech only. Similarly, from Table 11, random forest, neural networks and naïve Bayes all recorded some increases (average of 17.6%) in the validation with both speech and laughter while support vector machine and the logistic regression showed a slight decrease (average of 5.06%) in the performance metrics as compared to speech.

In addition, Table 12 shows an average of 3.25% improvement in the training and testing performance metrics of all the classifiers when both laughter and speech was used as compared to laughter only. Similarly, from Table 13, logistic regression showed a slight increase (average of 4.13%) in the validation with both speech and laughter while random forest, neural networks, naïve Bayes and support vector machine all recorded some decreases (average of 3.52%) in the performance metrics as compared to laughter only.

5.0 CONCLUSION

This paper presents the use of speech and laughter of people for person identification. Features were extracted (127 features in all) using the Librosa library in Python programming language via the Scientific Python Development Environment (SPYDER) IDE of the Anaconda software. These features were resampled in order to carter for imbalanced data distribution using the SMOTETomek library in Python. The resampled features were then exported to the Orange data mining software for further analysis. The information gain was used to rank the features so that the most important features were used for the analysis. The best performance was achieved with 80 features ranked by the information gain system and were used for the training, testing and validation. The neural network reported 92% classification accuracy, (0.91) F1_score, (0.92) precision and (0.82) recall. The high F1_score, precision and recall showed that the models did not overfit. The receiver operator characteristic (ROC) curve was closed to 1 for all the machine learning algorithms both for training and testing as well as validation. There was a slight improvement in the training and testing performance metrics of all the classifiers when both laughter and speech was used as compared to speech (an average of 1%) and laughter (an average of 3.25%) independently. Similarly, Random Forest, Neural Networks and Naïve Bayes all recorded some increases (average of 17.6%) in the validation with both speech and laughter while Support Vector Machine and the Logistic Regression showed a slight decrease (average of 5.06%) in the performance metrics as compared to speech. From the foregoing research, result showed that there is biometric trait in laughter which when combined with speech can improve the recognition rate. This research is useful for forensic application for example in criminal identification in conversation where people speak and laugh in between. Future studies will seek to improve the model through the acquisition of more volunteers for the dataset, and utilize recent developments in deep learning algorithm.

REFERENCES

Abd El-Moneim, S., Nassar, M. A., Dessouky, M. I., Ismail, N. A., El-Fishawy, A. S., Abd El-Samie, F. E. (2020). Text-independent speaker recognition using LSTM-RNN and speech enhancement. *Multimedia Tools and Applications*, 1-16.

Adekitan, A. I., Salau, O. (2020). Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. *SN Applied Sciences*, *2*(1), 8.

Arora, S. V., Vig, R. (2020). An efficient text-independent speaker verification for short utterance data from Mobile devices. *Multimedia Tools and Applications*, *79*(3), 3049-3074.

Assunção, G., Menezes, P., Perdigão, F. (2020). Speaker Awareness for Speech Emotion Recognition. *International Journal of Online and Biomedical Engineering (iJOE)*, *16*(04), 15-22.

Bäckström, Tom (2019). Introduction to Speech Processing. Online Blog, Retrieved from https://wiki.aalto.fi/pages/viewpage.action?pageId=149890776.

Bakmand-Mikalski, D., Rasmussen, A. H., Christensen, N. O. (2007). Speaker identification. *Master Thesis*. Retrieved from http://www2.compute.dtu.dk/pubdb/pubs/5580-full.html

Belin, P., Fecteau, S., Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129-135.

Bhogal, K. S., Hamilton, I. R. A., Kanevsky, D., Pickover, C. A., Sand, A. R. (2012). *U.S. Patent No. 8,140,340*. Washington, DC: U.S. Patent and Trademark Office.

Boersma, P., Weenink, D. (2015). Praat: doing phonetics by computer [Computer program, version 6.0. 06]. retrieved 18 December 2015 from http://www.praat.org.

Cakmak, H., Urbain, J., Tilmanne, J., Dutoit, T. (2014). Evaluation of HMM-based visual laughter synthesis. In *2014 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4578-4582). IEEE.

Chettri, B., Stoller, D., Morfi, V., Ramírez, M. A. M., Benetos, E., Sturm, B. L. (2019). Ensemble models for spoofing detection in automatic speaker verification. *arXiv preprint arXiv:1904.04589*.

Delac, K., Grgic, M. (2004, June). A survey of biometric recognition methods. In *Proceedings. Elmar-2004. 46th International Symposium on Electronics in Marine* (pp. 184-193). IEEE.

Devillers, L., Vidrascu, L. (2007). Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Interdisciplinary workshop on the phonetics of laughter* (p. 37).

El Ayadi, M., Hassan, A. K. S., Abdel-Naby, A., Elgendy, O. A. (2017). Text-independent speaker identification using robust statistics estimation. *Speech Communication*, *92*, 52-63.

Farzaneh, M., Toroghi, R. M. (2020). Robust Audio Watermarking Using Graph-based Transform and Singular Value Decomposition. *arXiv preprint arXiv:2003.05223*.

Fayek, H. (2016). Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between. *URL: https://haythamfayek. com/2016/04/21/speech-processingfor-machine-learning. html*.

Feng, L. (2004). *Speaker recognition* (Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark). Retrieved from https://www.researchgate.net/publication/259333765

Folorunso, C. O., Asaolu, O. S., Popoola, O. P. (2019). A Review of Voice-Base Person Identification: State-of-the-Art. *Covenant Journal of Engineering Technology*, (CJET) *3*(1). ISSN: p 2682-5317 e 2682-5325 DOI: 10.20370/2cdk-7y54. https://journals.covenantuniversity.edu.ng/index.php/cjet/article/view/1635/978

Folorunso, C.O., Asaolu, O.S. and Popoola, O.P. (2020) 'Laughter signature: a novel biometric trait for person identification', *Int. J. Biometrics*, Vol. 12, No. 3, pp.283–300.

Fromont, L. A., Royle, P., Steinhauer, K. (2020). Growing Random Forests reveals that exposure and proficiency best account for individual variability in L2 (and L1) brain potentials for syntax and semantics. *Brain and Language*, *204*, 104770.

Gyanendra K. Verma "Multi-feature Fusion for Closed Set Text Independent Speaker Identification" International conference on information intelligence, systems, technology and management, Springer (2011), pp. 170-179

Gosztolya, G., Beke, A., Neuberger, T., Tóth, L. (2016). Laughter classification using Deep Rectifier Neural Networks with a minimal feature subset. *Archives of Acoustics*, *41*.

Jagdale, S. M., Shinde, A. A., Chitode, J. S. (2020). Robust Speaker Recognition Based on Low-Level-and Prosodic-Level-Features. In *Advances in Data Sciences, Security and Applications* (pp. 267-274). Springer, Singapore.

Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Ali, I. (2020). Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network. *IEEE Access*, *8*, 32187-32202.

Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R. K. (2020). Speech Emotion Recognition using Support Vector Machine. *arXiv preprint arXiv:2002.07590*.

Jason Brownlee (2020). How to Combine Oversampling and Undersampling for Imbalanced Classification. Online Machine Learning Course, Retrieved from https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification/ on May 7, 2020.

Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., Cai, L. H. (2002). Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo* (Vol. 1, pp. 113-116). IEEE.

Kamble, M. R., Sailor, H. B., Patil, H. A., Li, H. (2020). Advances in anti-spoofing: from the perspective of ASVspoof challenges. *APSIPA Transactions on Signal and Information Processing*, *9*.

Kattel, M., Nepal, A., Shah, A. K., Shrestha, D. (2019). Chroma feature extraction. In *Conference: Chroma Feature Extraction using Fourier Transform*.

Kaur, J., Kaur, S. (2016). A Brief Review: Voice Biometric for Speaker Verification in Attendance Systems. *Imp. J. Interdiscip. Res*, *2*(10).

Kinnunen, T., Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*(1), 12-40.

Korshunov, P., Marcel, S. (2016). Joint operation of voice biometrics and presentation attack detection. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (pp. 1-6). IEEE.

Krawczyk, S., Jain, A. K. (2005). Securing electronic medical records using biometric authentication. In *International Conference on Audio-and Video-Based Biometric Person Authentication* (pp. 1110-1119). Springer, Berlin, Heidelberg.

Laskowski, K., Schultz, T. (2008). Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. In *International Workshop on Machine Learning for Multimodal Interaction* (pp. 149-160). Springer, Berlin, Heidelberg.

Levitan, S. I., Mishra, T., Bangalore, S. (2016). Automatic identification of gender from speech. In *Proceeding of speech prosody* (pp. 84-88).

Low, D. M., Bentley, K. H., Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96-116.

Mazaira-Fernandez, L. M., Álvarez-Marquina, A., Gómez-Vilda, P. (2015). Improving speaker recognition by biometric voice deconstruction. *Frontiers in bioengineering and biotechnology*, *3*, 126.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18-25).

Medikonda, J., Bhardwaj, S., Madasu, H. (2020). An information set-based robust text-independent speaker authentication. *Soft Computing*, *24*(7), 5271-5287.

Mokgonyane, T. B., Sefara, T. J., Modipa, T. I., Mogale, M. M., Manamela, M. J., Manamela, P. J. (2019). Automatic speaker recognition system based on machine learning algorithms. In *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)* (pp. 141-146). IEEE.

Nagels, L., Gaudrain, E., Vickers, D., Hendriks, P., Başkent, D. (2020). Development of voice perception is dissociated across gender cues in school-age children. *Scientific Reports*, *10*(1), 1-11.

Nagakrishnan, R., Revathi, A. (2020). A robust cryptosystem to enhance the security in speech-based person authentication. *Multimedia Tools and Applications*. Retrieved from https://doi.org/10.1007/s11042-020-08846-1

Palo, H. K., Behera, D., Rout, B. C. (2020). Comparison of Classifiers for Speech Emotion Recognition (SER) with Discriminative Spectral Features. In *Advances in Intelligent Computing and Communication* (pp. 78-85). Springer, Singapore.

ResearchGate (2014). How to estimate a person's vocal tract length, using a recorded audio file. Online Blog Retrieved from https://www.researchgate.net/post/How_can_I_estimate_a_persons_vocal_tract_length_using_a_recorded_audio_file#:~:text=I'm%20aware%20of%20the,to%20an%20unconstricted%20vocal%20tract.

Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matějka, P., Burget, L., Glembek, O. (2020). End-to-end DNN based text-independent speaker recognition for long and short utterances. *Computer Speech & Language*, *59*, 22-35.

Ruch, W., Wagner, L., Hofmann, J. (2019). A lexical approach to laughter classification: Natural language distinguishes six (classes of) formal characteristics. *Current Psychology*, 1-13. https://doi.org/10.1007/s12144-019-00369-9

Scheffer, N., Ferrer, L., Lawson, A., Lei, Y., McLaren, M. (2013). Recent developments in voice biometrics: Robustness and high accuracy. In *2013 IEEE International Conference on Technologies for Homeland Security (HST)* (pp. 447-452). IEEE.

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, *103*(1), 588-601.

Shakil, K. A., Zareen, F. J., Alam, M., Jabin, S. (2020). BAM Health Cloud: A biometric authentication and data management system for healthcare data in cloud. *Journal of King Saud University-Computer and Information Sciences*, *32*(1), 57-64.

Sheikh, I. A., Kopparapu, S. K., Vachhani, B. B., Garlapati, B. M., Chalamala, S. R. (2020). Method and system for muting classified information from an audio. *U.S. Patent Application No. 16/254,387*.

Singh, A., Joshi, A. M. (2020). Speaker Identification Through Natural and Whisper Speech Signal. In *Optical and Wireless Technologies* (pp. 223-231). Springer, Singapore.

Sun, L., Gu, T., Xie, K., Chen, J. (2019). Text-independent speaker identification based on deep Gaussian correlation supervector. *International Journal of Speech Technology*, *22*(2), 449-457.

Tawara, N., Ogawa, A., Iwata, T., Delcroix, M., Ogawa, T. (2020). Frame-Level Phoneme-Invariant Speaker Embedding for Text-Independent Speaker Recognition on Extremely Short Utterances. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6799-6803). IEEE.

Thoman, C. (2009). *Model-Based Classification of Speech Audio*. Florida Atlantic University. Retrieved from http://fau.digital.flvc.org/islandora/object/fau%3A3410/datastream/OBJ/view/Model-based_classification_of_speech_audio.pdf.

Tran, D., Wagner, M., Lau, Y. W., Gen, M. (2004). Fuzzy methods for voice-based person authentication. *IEEJ Transactions on Electronics, Information and Systems*, *124*(10), 1958-1963.

Urbain, J., Dutoit, T. (2011, October). A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 397-406). Springer, Berlin, Heidelberg.

Urone, Paul Peter and Hinrichs, Roger (2020). College Physics. Online course published by OpenStax. Retrieved from https://openstax.org/books/college-physics/pages/16-11-energy-in-waves-intensity.

Vatsa, M., Singh, R., Noore, A. (2009). Feature based RDWT watermarking for multimodal biometric system. *Image and Vision Computing*, *27*(3), 293-304.

Yasmin, G., Dhara, S., Mahindar, R., Das, A. K. (2019). Speaker Identification from Mixture of Speech and Non-speech Audio Signal. In *Soft Computing in Data Analytics* (pp. 473-482). Springer, Singapore.