Development of Children Gender Classification System Using Speech

Khin Aye Chan [#1], Su Su Maung [#2], Khine Thin Zar [#3]
[#]*Department of Computer Engineering and Information Technology*
Yangon Technological University
Yangon, Republic of the Union of Myanmar
[1] khinayechanchan@gmail.com
[2] susuela@gmail.com
[3] ngengeko@gmail.com

Abstract- Today's classification of gender is one of the most important procedures in speech processing. Gender identification task from children's speech is a challenging problem as there's no significant difference in the acoustic properties of male and female children. In this thesis, children gender classification system makes investigation on the efficient features to discriminate the gender from children's speech. The Mel Frequency Cepstral Coefficient (MFCC) method is used for extracting features from speech signals. Features are evaluated using nonlinear classifiers, namely Artificial Neural Network (ANN), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB).This gender classification system is implemented by applying Python programming language and the experimental results has been analysed. From the results RF classifier achieves the highest accuracy of 83% and performs better compared with other machine learning algorithms to classify the gender of a child using MFCC features of voice.

Keywords – Speech Processing, Gender Classification, Feature Extraction, MFCC, RF

I. INTRODUCTION

Speech recognition is the knowledge base subfield of linguistics that develops methodologies and technologies that permit the recognition and translation of spoken communication into text by computers. It incorporates data and analysis within the linguistics, computer science, and electrical engineering fields. Some speech recognition systems need "training" wherever a private speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the identification of that person's speech, leading to inflated accuracy. Systems that don't use training are referred to as "speaker independent" systems. Systems that use training are referred to as "speaker dependent" [1].

Automatic gender classification from speech is widely applied in speech recognition. Many applications including speaker identification, speaker segmentation, smart human computer interaction, biometrics social robots, and audios or videos content indexing, etc., use gender classification. Gender identification can improve the prediction of other speaker traits such as age and emotion, by jointly modeling gender with age (or emotion) or either by together modeling gender with age (or emotion) or during a pipelined manner. Generally humans can easily identify a person's age, gender and emotion by hearing to this person's voice. In some circumstances like conversations over the telephone, the genders of adults are easy to identify, but the genders of children are difficult to identify. Children generally have higher elementary and format frequencies than those of adults, thanks to a shorter vocal tract, smaller vocal folds, developing

articulators (e.g. tongue size and movement). Therefore, there is no significant difference in their acoustic-phonetic properties in both male and female child. This gender classification system examines the features efficient in discriminating the gender from children's speech with the help of feature extraction algorithms. The extraction of the relevant and important information from the speech signals of the human voice is an important task to produce a latter recognition performance. And then the system makes performance analysis about classification by using different ML algorithms [2].

In computer science, ML identifies to a type of data analysis that applies algorithms that learn from data. It is a type of artificial intelligence (AI) that provides systems with the ability to learn without being especially programmed. This implements computers to find data within data without human intervention. The main importance to know about ML is that data is being used to make predictions, not code. Data is dynamic so ML permits the system to learn and evolve with experience and the more data that is analyzed [3].

In this study, children gender classification system is proposed to solve the problem that gender identification of children is difficult than adults; it is confusing to identify whether the speaking child is male or female by combining ASR, feature extraction and ML algorithms. The proposed system is implemented by applying Python programming language and the experimental result has been analyzed.

II. DATA COLLECTION

The database was designed to create a training set of speech from children of KG to Grade V (age range 6 to 11 years). The database contains audio recordings of individuals from different grades. For each grade, 10 boys and 10 girls are selected. They speak 10 Myanmar sentences which are recorded in quiet classrooms. These sentences are mentioned below. Utterances from children are recorded with SONY Digital Stereo High Definition. These voice clips are preprocessed and evaluated. There are total of 1200 audio records. The female records contain 600 samples where male records have 600 samples. The prepared Myanmar sentences are:

1. မင်္ဂလာပါ။
2. ကျေးဇူးတင်ပါတယ်။
3. နာမည်ဘယ်လိုခေါ်လဲ။
4. ဘယ်သွားမလို့လဲ။
5. နေကောင်းလား။
6. ဘယ်မှာနေပါသလဲ။
7. စာမေးပွဲဖြေနိုင်လား။
8. အမေကိုခေါ်လာခဲ့ပါ။
9. ဘာဖြစ်လို့လဲ။
10. ဘာဝါသနာပါလဲ။

III. METHODOLOGY

A. Preprocessing

The speech recordings of children consist of many silence and unvoiced regions. Audio trimming is done to remove silence duration and unvoiced regions from start and end of each audio file. Librosa's functions are applied for preprocessing and librosa is a python package for music and audio analysis. The librosa.effects.trim function is used for trimming silence and unvoiced region. This function includes important parameters: input audio signal, threshold value to assign as silence, the number of samples per analysis frame and the number of samples between analysis frame.

B. Feature Extraction from Children's Speech

Feature extraction from a given signal is the most significant phase in gender identification. The most widely used features for speech recognition are the acoustic features, namely Mel Frequency Cepstral Coefficient (MFCC). In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound. The power spectrum describes the distribution of power into frequency components composing that signal. MFCCs are coefficients that collectively make up an MFC. The reason for MFCC being most commonly used for extracting features is that it is most nearest to the actual human auditory speech perception. It is based on the short term analysis, and thus from each frame of speech signal a MFCC vector is computed. MFCC feature extraction method is less complex in implementation and more effective and robust under various conditions. It is a standard method for feature extraction in speech recognition [4].

C. Classification

Artificial Neural Network (ANN) is a simple mathematical model of the brain which is used to process nonlinear relationships between inputs and outputs in parallel like a human brain does every second. Information flows through a neural network in two different ways. When the model is learning (being trained) or operating normally (after being trained either being used or tested), patterns of information from the dataset are being fed into the network via the input neurons, which trigger the layers of hidden neurons, and these in turn arrive at the output neurons. This is called a feedforward network. Each neuron receives inputs from the neurons to its left, and the inputs are multiplied by the weights of the connections they travel along. Feedforward neural network is considered for this experimentation. These networks are mostly used for supervised machine learning tasks where the target function is already known [5].

Random Forest (RF) consists of a large number of individual decision trees that operate as an ensemble. Ensemble is a data mining technique composed of number of individual classifiers to classify the data to generate new instances of data. RF is the most popular ensemble technique that can use both for classification and regression kinds of problems. Each individual tree in the RF spits out a class prediction and the class with the most votes becomes our model's prediction. RF uses decision trees as base classifier. Decision tree breaks down a data set into smaller and smaller subsets while at the same

time an associated decision tree is incrementally developed. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data [6].

Logistic Regression (LR) called the logistic model or logit model, analyzes the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event such as pass/fail, win/lose, alive/dead or healthy/sick by fitting data to a logistic curve. There are two models of LR, binary LR and multinomial LR. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinomial logistic regression can be employed [7].

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. When data are unlabeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups [8].

A Gaussian Naive Bayes classifier (GNB) is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem. A GNB algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution. Primarily NB is a linear classifier, which is a supervised machine learning method and works as a probabilistic classifier as well. GNB is the simplest and the most popular one. When handling real-time data with continuous distribution, NB classifier considers that the big data is generated through a Gaussian process with normal distribution. In general the Naive Bayes classifier is not linear, but if the likelihood factors are from exponential families, the Naive Bayes classifier corresponds to a linear classifier in a particular feature space. GNB is an algorithm having a probabilistic approach. It involves prior and posterior probability calculation of the classes in the dataset and the test data given a class respectively [9].

IV. RESULTS AND DISCUSSION

Evaluating machine learning algorithm is an essential part of this study. The performance of the proposed system is specified based on classification results for training and testing data set. Two testing is done to estimate the performance of the machine learning classification models: simple train-test split and k-fold cross validation method. Different performance metrics are also used to evaluate performance of different machine learning

algorithms. Performance metrics used in this study are confusion matrix, precision, recall, F1-score, support are used to evaluate performance of classifiers.

A. Simple Train-test Split

Splitting test dataset and train dataset can make by any ratio and size of datasets can be declared as test size and train size. In the beginning, all of the stated classifier models were trained and tested for 70% and 30% of the total dataset respectively. Then in different ratios of the experiment it was changed to 80% with 20% and 90% and 10%. Classification results of individual process are described in TABLE I.

TABLE I. ACCURACY RESULTS

Data Ratio	**70:30**		**80:20**		**90:10**	
Classifiers	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
RF	99%	84%	99%	81%	99%	84%
ANN	83%	81%	80%	80%	82%	76%
SVM	76%	76%	76%	73%	76%	75%
LR	75%	76%	77%	68%	77%	76%
GNB	77%	76%	77%	75%	76%	76%

According to resulting accuracies, it can be seen that train-test ratio 70: 30 achieves highest accuracy for testing. RF classifier gets 84% of accuracy and has better performance than other classifiers in 70:30 data ratio. ANN with 81% is low accuracy compared to the RF. SVM, LR and GNB have an accuracy result of around 76%, which was fairly satisfying but when the amount of test size 20 is changed, the accuracy is slightly fall.

B. K-Fold Cross-Validation

K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. This procedure is implemented by randomly dividing the set of observations into k groups, or folds, of approximately equal size. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. TABLE II describes average testing accuracy of classifiers by 10-fold cross validation.

RF is efficient in building an accurate classifier which can efficiently run on the small and large sized datasets of non-linear nature. Hence RF is observed achieving good accuracy compared to the other four classifiers. ANN is low accuracy compared to the RF and that is 76%. Though ANN is efficient in modeling the non-linear data, small size of data set may affect the performance of ANN as they need large data for training. RF is efficient in discriminating features non-linear in nature. It also works well with the small

sized data. RF outperforms ANN with overall highest accuracy of 84% for feature dataset used in this system. Moreover, RF classifier achieves 83 % for male prediction and 86 % for female prediction.

SVM is a largely used model to deal with two class identifier .This model is trained and tested for a increase in accuracy but it has only 75% average accuracy. The implementation is simple and efficient but the score was not satisfactory.LR and GNB are tested for the same amount of training and testing data. They also get around 75% average accuracy as train-test split testing.

TABLE II. CROSS VALIDATION SCORE

Classifiers	RF	ANN	SVM	LR	GNB
10-fold Cross Validation Score	80%	77%	75%	73%	75%
	80%	76%	70%	69%	71%
	88%	75%	77%	76%	78%
	85%	80%	74%	75%	77%
	79%	77%	78%	79%	75%
	85%	76%	76%	77%	78%
	75%	75%	71%	74%	68%
	81%	74%	79%	80%	81%
	85%	78%	77%	78%	77%
	89%	72%	68%	68%	82%
Average	83%	76%	75%	75%	76%

C. Performance Matrices

The performance metrics chosen to evaluate machine learning models are very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. Metrics used for evaluation of children gender classification are precision, recall, F1-score and support. Precision is the amount of positive predictions that were correct. Precision is the number of correct positive results divided by the number of positive results predicted by the classifier. Recall refers to the percentage of total relevant results correctly classified by the algorithm. In other words, recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). F1 Score is the

Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It shows how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). TABLE III gives performance measures of each classifiers calculated on feature dataset divided into 70% training and 30% testing.

TABLE III. PERFORMANCE MATRICES

Classifiers	Gender	Precision	Recall	F1-score
RF	Female	83%	86%	84%
	Male	85%	83%	84%
ANN	Female	81%	79%	80%
	Male	80%	82%	81%
LR	Female	77%	77%	77%
	Male	77%	77%	77%
SVM	Female	78%	74%	76%
	Male	75%	79%	77%
GNB	Female	76%	78%	77%
	Male	78%	75%	76%

IV. CONCLUSION

This study presents implementation of children gender classification using speech. In the proposed system, voice feature extraction, machine learning (ML) and classification algorithms are combined. The system conducts a number of experiments with MFCC features dataset with the goal of developing a children's gender recognition system. The gender classification system is implemented by applying Python programming language and the experimental results have been analyzed. This system can classifies voice of children between 6 to 11 years of age with average accuracy of 83%. Based on the nonlinear nature of the data, classifiers efficient in discriminating non-linear data, namely RF, ANN, SVM, LR and GNB are considered. The RF outperforms the other classifiers with an average accuracy of 83% for gender classification. Classifiers can identify well for age group 9-11 years and most of misclassifications occur in age group 6-8 years. In train-test split testing, the best percentage of divided training data and testing data in this system is 70:30. RF classifier achieves the better classification results than other

classifiers by using this test size. Testing accuracy of RF model is 84%. According to each gender, RF classifier achieves 83 % for male prediction and 86 % for female prediction. For cross validation testing, the average accuracy of RF classifier is 83%. The analysis of the results shows that the performance of the proposed system is good, as the average accuracy of the best classifier is 83%.

ACKNOWLEDGMENTS

The author would like to express her deepest thanks and gratitude to Dr. Nyein Nyein Oo, Professor and Head of the Department of Computer Engineering and Information Technology of the Yangon Technological University. The author's special thanks go to her supervisor Dr. Su Su Maung, Associate Professor of the Department of Computer Engineering and Information Technology of the Yangon Technological University, for her invaluable advice and suggestion throughout the study. The author would like to express her thanks to her co-supervisor Dr. Khine Thin Zar, for her valuable comments and guidance during this study. Finally, her special thanks go to all who help her with necessary assistance for this study.

REFERENCES

[1] Yoshida,K. 1991. "Speech Recognition System". *Journal of the Japan Society for Precision Engineering* 57, no.11: 1924-1927.

[2] Raahul, A., Sapthagiri, R.,and Pankaj, K et al. 2017. "Voice based gender classification using machine learning". *IOP Conference Series: Materials Science and Engineering* 263, no.4 (March):0-9.Bastanlar,Y., and Ozuysal,M. 2014. "Introduction to Machine Learning". *Methods in Molecular Biology* 1107 (October): 105-128.

[3] Zupan, J. 1994. "The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition". *International Science Index, Computer and Informatics Engineering* 11, no.1 (June): 1100-1104.

[4] Fawaz,S.A., and Dia, A. 2017. "Introduction to Artificial Neural Network (ANN) Methods: What They are and How to Use Them". *Acta Chimica Slovenica* 41 (September): 327-327.

[5] Cutler,D., Edwards, T., Beard, K et al. 2007. "Random Forests For Classification In Ecology". *Ecology* 88, no.11 (November): 2783-2792.

[6] Sperandei, S. 2014. "Understanding logistic regression analysis". *Biochemical and Medica1* 24, no.1(February): 12-18.

[7] Hsu,C., Chang, C., and Lin, C. 2008. "A Practical Guide to Support Vector Classification". *BJU international* 101. no.1:1396-1400.

[8] Eberhardt, L., and Breiwick, J. 2013. "Learning the Naive Bayes Classifier with Optimization Models". International Journal of Applied Mathematics and Computer Science 23, no.4 (December): 787-795.