Speech Command Classification System for Sinhala Language based on Automatic Speech Recognition

Thilini Dinushika, Lakshika Kavmini, Pamoda Abeyawardhana, Uthayasanker Thayasivam and Sanath Jayasena
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
Email: dinushiranagalage.15@cse.mrt.ac.lk

***Abstract*—Conversational Artificial Intelligence is revolutionizing the world with its power of converting the conventional computer to a human-like-computer. Exploiting the speaker's intention is one of the major aspects in the field of conversational Artificial Intelligence. A significant challenge that hinders the effectiveness of identifying the speaker's intention is the lack of language resources. To address this issue, we present a domain-specific speech command classification system for Sinhala, a low-resourced language. It accomplishes intent detection for the spoken Sinhala language using Automatic Speech Recognition and Natural Language Understanding. The proposed system can be effectively utilized in value-added applications such as Sinhala speech dialog systems. The system consists of an Automatic Speech Recognition engine to convert continuous natural human voice in Sinhala language to its textual representation and a text classifier to accurately understand the user intention. We also present a novel dataset for this task, 4.15 hours of Sinhala speech corpus in the banking domain. Our new Sinhala speech command classification system provides an accuracy of 89.7% in predicting the intent of an utterance. It outperforms the state-of-the-art direct speech-to-intent classification systems developed for the Sinhala language. Moreover, the Automatic Speech Recognition engine shows the Word Error Rate as 12.04% and the Sentence Error Rate as 21.56%. In addition, our experiments provide useful insights on speech-to-intent classification to researchers in low resource spoken language understanding.**

Keywords: Sinhala Speech Command Classification, Automatic Speech Recognition, Intent Classification

I. Introduction

Recent advances in conversational Artificial Intelligence (AI) have resulted in conversation-based applications with a wide range of supported platforms. Google Assistant [1] and Amazon Alexa [2] are two such prominent commercial conversational agents that assist in voice-based control ranging from smartphones to home automation. Speech command classification, also known as *intent classification*, is a key research area in the field of conversational AI. Yet obtaining state-of-the-art results in classifying free-form speech commands of low-resourced languages is challenging [3].

Sinhala, an Indo-Aryan language and the official language of Sri Lanka is recognized as a low-resourced language [4]. Developing an accurate Sinhala speech command classification system can be considered as an initiative to promote Sinhala community to reach the digitized world through their native spoken language. But the lack of resources for Sinhala language hinders the development of accurate models for intent classification.

The combination of Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) is known for its high level of accuracy in speech command classification of resource full languages. In contrast, Buddhika et al. [b3] and Karunanayake et al. [5] claim that the resource limitation in Sinhala causes suboptimal performance in both the ASR and NLU. Hence, they have proposed a direct classification approach for Sinhala speech-to-intent. To the best of our knowledge, these were the only prior efforts of Sinhala speech to intent mapping. Other Sinhala speech related efforts have focused only on speech-to-text conversion, speech classification or speech clustering.

In this paper, we present the first domain-specific speech command classification system for Sinhala language using ASR and text classification. It classifies banking domain related free-form Sinhala speech commands to their respective intents. Our novel approach aims at exploiting a fine-tuned ASR to gain better accuracy in classification and outperforms the state-of-the-art direct speech-to-intent mapping. Further, we analyze the relevant gain in using a fine-tuned ASR as opposed to the direct classification approach and associated efforts. Though we focus the Sinhala language in this research, the insights presented through this analysis are applicable to other low-resourced language research for developing speech-to-intent systems efficiently and effectively.

The contributions of this paper are: 1. We present a new Sinhala speech corpus in the banking domain 2. We present a novel continuous speech recognition system for Sinhala using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) 3. We demonstrate our Sinhala speech command classification system based on ASR and evaluate its performance.

The rest of the paper is organized as follows. Section two presents the related work, and section three includes the background. Section four and five present the dataset and the methodology, respectively. Section six presents our experiments and section seven contains a comparative discussion about the results. Finally, section eight carries our conclusion and future work.

II. Related Work

Buddhika et al. [3] have presented a Sinhala speech command classification system which uses a direct speech-to-intent mapping without an intermediary (e.g., text) representation.

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

They have implemented a classification algorithm based on Feed Forward Neural Networks to classify Sinhala speech commands. Mel-frequency Cepstral Coefcients (MFCC) have been extracted from speech signals as feature vectors. Using only 10 hours of speech data, they have achieved a classification accuracy of 74%. In addition, Karunanayake et al. [5] have extended the same direct classification approach together with transfer learning to classify domain specific free-form Sinhala and Tamil speech commands. The system has been developed by utilizing a character probability map from an ASR model trained on the English language and has achieved a reasonable accuracy for both Sinhala and Tamil data sets. In contrast to this, our effort employs an intermediary text representation to outperform this direct approach of speech command classification.

Most previous research on speech command classification of both popular and low-resourced spoken languages have been dominated by the combination of ASR and NLU components. An ASR module converts the speech into a text representation and this text is fed to a text classifier in order to understand the intention of the speaker [6] [7]. As shown by Rao et al. [6], this cascading approach can be effectively utilized for intent classification by developing high accurate, fine-tuned ASR systems. Hence, our research focuses on evaluating the applicability of ASR and NLU techniques for Sinhala speech command classification.

The advancements in ASR techniques have shown proven performance in recognizing the speech of several languages. A review on applying HMM for speech recognition [8] has suggested that HMM is flexible and easily implemented for speech recognition of any language due to its inherent mathematical framework. Thus, several speech recognition toolkits that support HMM are widely available [9] [10]. Zissman et al. [11] have presented the combination of GMM together with HMM as an approach to further enhance the accuracy of speech recognition.

There are few prior researches on Sinhala ASR systems implemented using HMMs. The Interactive Voice Response (IVR) system presented by Manamperi et al. [12] shows a Word Error Rate (WER) of 11.2% for digit recognition and a Sentence Error Rate (SER) of 5.7% for song recognition. Also, Nadungodage et al. [13] have achieved an accuracy of 75.74% in recognizing continuous Sinhala speech commands using a data set collected from the voice of a single speaker. In addition, [14] [15] can be identified as successful attempts to recognize isolated Sinhala words with reasonable accuracy. We refer these approaches to build the language model and phoneme dictionary of our ASR system. In addition, ASR systems [16] and [17], for Bengali and Oriya languages, respectively, similar languages to the spoken Sinhala language, can be taken as reference models to develop a fine-tuned ASR system for Sinhala.

III. Background

This section carries a brief overview of the major techniques we employed to develop the ASR engine and the text classifier of the proposed Sinhala speech command classification system.

A. MFCCs for feature extraction

This is a method of extracting frequency information in speech signals and converting them into coefficients. Since MFCCs simulate the properties of human auditory system, they are widely applied in speech processing [18].

B. GMM-HMM for acoustic modeling

HMM is a statistical model to compute the probability of a sequence of feature vector observations based on some sequence of hidden state transitions. GMM is combined with HMM to model the distribution of real-valued feature vectors corresponding to each HMM state. This combination of GMM-HMM is widely adopted for acoustic modeling in ASR systems [9].

C. N-gram model

The N-gram model implementation predicts the posterior probability of observing a word (W_n) in a sequence of words (i.e a sentence), given that the words W_1, W_2, $\ldots...W_{n-1}$ are preceding in the sequence i.e. $P(W_n|W_1...W_{n-1})$. (1) calculates this posterior probability using the word counts in a given language corpus.

$$P(W_n|W_1...W_{n-1}) = \frac{count(W_1, \ldots, W_{n-1}, W_n)}{count(W_1, .., W_{n-1})} \quad (1)$$

D. Support Vector Machines

Support Vector Machines (SVM) is a binary classification algorithm that determines the decision boundary between feature vectors of two classes. It can be scaled well for multi-class classification and can be generalized well in high dimensional feature spaces. SVM is recognized as an easy-to-use and robust technique for text classification [19].

IV. Sinhala Speech Corpus

The scope of existing Sinhala speech corpora are not extensible to model a conversation between a human and an agent. Thus, a novel Sinhala speech corpus is built with the intention of developing a Sinhala speech dialog system. We selected the banking domain, analyzed few conversations between a customer and a customer service assistant of a bank during the process of opening a new bank account. We identified 14 basic intentions a customer would express during this common conversation.

First, we used a crowdsourcing approach to identify different inflections on how each of the intents would be uttered in spoken Sinhala language. A Google form including the 14 predefined intents was distributed among 130 participants covering different age groups. We requested them to provide alternative ways in which people express each of these intents in spoken Sinhala language. In addition, participants were selected to capture the different dialects of the Sinhala language spoken in different regions of the country.

TABLE I: Few inflections under a sample intent

Intent		Different inflections (In English transliteration)
Sinhala intent in English transliteration	*Meaning*	
nawa giNumak wiwurta kiriema	Request to open a new bank account	maTa nawa giNumak wiwurta karanna ona
		maTa alut giNumak wiwurta karanna ona
		giNumak arinna ona

The data was analyzed and a finalized set of inflections corresponding to each intent was created with the help of few language experts. Table 1 includes an example of few inflections identified under the intent - "Request to open a new bank account" in English transliteration.

Voicer, a web/smartphone based crowdsourcing tool presented in [20]: was used to collect speech samples. The tool was re-configured to capture balanced amounts of speech clips for each intent. Multiple users can simultaneously access the tool and record their voice by uttering inflection commands prompted by the tool. The data collection process was conducted under uncontrolled environmental conditions.

The data was collected from 120 speakers representing 60% males and 40% females. 30% of total speakers were university students and the rest from the general community within the age group of 25 to 60 years. The average length of an individual recording ranges from one to three seconds.

Using Voicer, we collected a total of 9650 speech clips for all infections under the 14 intents. These speech clips were validated manually and subjected to noise removal. After removing all flawed clips with over recordings, halfway-stopped, and high noise profiles, 8977 speech clips were shortlisted to build the corpus. The final Sinhala speech corpus was 4.15 hours long. The corpus was divided into training and testing set with 80%, 20% ratio respectively.

V. Methodology

The proposed Sinhala speech command classification system is designed with two main subcomponents: a domain specific ASR engine and a text classifier. The ASR engine is developed using a combination of GMM-HMM to convert continuous Sinhala speech into text. The text classifier is modeled using SVM and it predicts the intent of the text-output generated by the ASR. The high-level architecture of the system is depicted in Fig.1. The rest of the section describes the implementation and training process of each component.

A. ASR engine

The ASR engine is developed using a statistical approach as a combination of three basic models: acoustic model, language model and the dictionary model. A 3.32 hours worth speech corpus is used to train the acoustic model. During the training, MFCCs are extracted as the acoustic feature vectors in order to characterize speech signals. The following subsections describe the process of building the aforementioned models during the training phase.

1) Acoustic model: The acoustic model of the ASR is designed to capture important attributes from the extracted feature vectors of speech signals. To build the acoustic model, the system is fed with audio clips and their respective text transcriptions. The GMM-HMM combination is used to capture the sequence of phones for acoustic modeling. The procedure followed to build the acoustic model is illustrated in Fig. 2.

During the acoustic model training, separate HMM models are trained for each phoneme (smallest sub-word unit in speech) identified within the speech corpus. The topology of each HMM model is assigned with three hidden states, which the first state for the transition into the phoneme, second for the middle part and the third state for the transition out of the phoneme. This phoneme-based HMM modeling is known as monophone training. The monophone models do not capture any contextual information variants based on preceding or following phonemes.

Next, these monophone models are chained together based on the phoneme constitution of words in order to form new models to recognize words. These context-dependent models which are referred to as triphones, reflect the contextual variations in phoneme occurrences within a speech wave. Finally, a network of HMM models is trained by concatenating these matching triphone models in order to recognize words in continuous speech.

For every state in an HMM, a GMM model is trained. This probabilistic model is built using the Expectation-

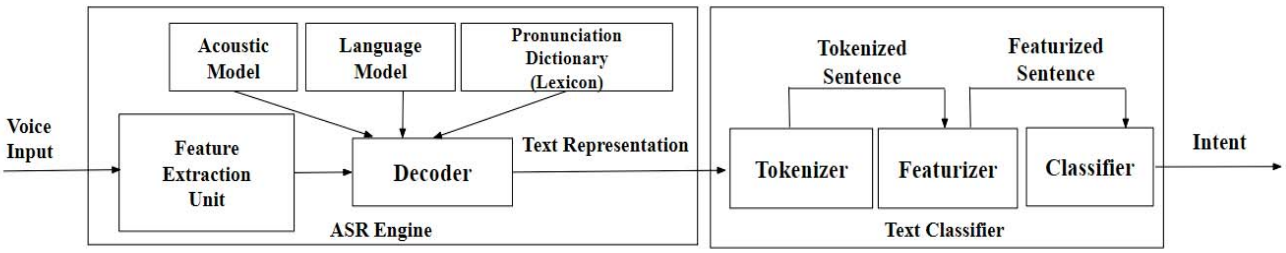

Fig. 1: Architecture of the speech command classification system

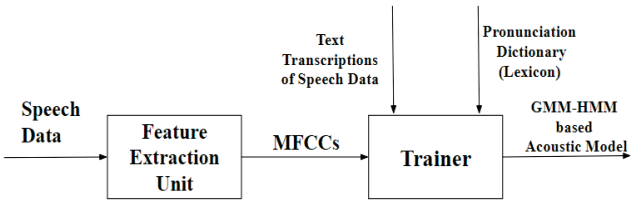

Fig. 2: The process of training the acoustic model

Maximization algorithm. These GMM models are responsible for modeling the emission probability of each state, the probability of each HMM state's association to a speech feature distribution.

2) Language model: The language model essentially captures the grammar of the language. Since our ASR system is based on a limited domain of vocabulary, the N-gram model is chosen instead of hardcoded grammar rules. To train the language model, we have implemented 3-Gram model with the help of SRI Language Modeling Toolkit (SRILM).

3) Dictionary model: A phoneme is the smallest unit of speech which differentiates a word from another. The phonetic dictionary model is created manually by including the phoneme representation of all words of the selected vocabulary. ISO 15919 - Transliteration of Devanagari and related Indic scripts into Latin characters [21] was referred to construct the phonetic representation of words.

The Sinhala language has 47 phonemes comprised of 14 vowels and 33 consonants. Our domain specific system based on the banking domain has 96 Sinhala words in its dictionary.

B. Text classifier

The textual representation of user utterance generated by the ASR engine is fed into the text classifier in order to predict the most relevant intent out of the 14 pre-defined intent classes. The Sinhala text corpus comprised of infections under 14 intents which was collected through the crowdsourcing, is used to train the classifier models.

During the training, the sentences in the text data corpus is directly fed to the tokenizer in raw format without subjecting to any text normalization. The tokenizer breaks the sentences into tokens (i.e. words) and passes an array of tokens to the featurizer. The featurization component builds a dictionary of words including all the unique words within the text corpus (in our case 96 different words). It assigns a unique integer to each of the words and computes a score for each word using Term Frequency and Inverse Document Frequency (TF-IDF) weighting scheme. Next, each sentence in the corpus is given a vector representation based on the created dictionary. These labeled feature vectors are utilized to train a multi-class text classifier based on SVM while incorporating 'one vs all' approach [22].

VI. Experiments

The decoder of the ASR utilizes the trained acoustic model and the language model to convert a given speech sample to its text representation. This text-output is fed into the text classifier to predict the intent and evaluate the intent classification accuracy.

The performance of the ASR is evaluated using both the WER and SER. The WER is the number of incorrectly identified words out of the total number of words in the test speech corpus. SER is the number of incorrectly identified sentences out of the total number of sentences in the same test speech corpus.

The domain-specific intent classification system presented by Buddhika et al. [3] is selected as our benchmark. It has shown a 74% of test accuracy for intent detection using its direct approach with 10 hours of speech data. As the final outcome, our Sinhala speech command classification system achieves an overall intent classification accuracy of 89.7% only using a 4.15 hours speech data corpus with its GMM-HMM based acoustic model in the ASR and SVM as the text classifier.

Our ASR engine based on GMM-HMM shows a WER of 12.04% and an SER of 21.56%. In addition, the variation of WER and SER with respect to the duration of training speech data corpus is presented in Table 3.

A comparative analysis is conducted to understand the impact of text classification technique on the overall intent classification accuracy. Accordingly, different text classification models are trained using, 1. Neural Network Embeddings 2. Naive Bayesian 3. Logistic Regression and used in place of the SVM classifier [23] [24]. Table 4 presents the accuracy of the speech command classification system against each of the classifier models.

VII. Discussion

The results show that our proposed methodology of utilizing a fine-tuned ASR module outperforms the state-of-the-art direct speech-to-intent mapping. Hence, we prove that utilizing an ASR and following through an intermediary stage of text representation is a viable solution for the challenge of recognizing low-resourced spoken languages. Moreover, our ASR based speech command classification system shows a significant accuracy even with half-sized data corpus compared to [3].

TABLE II: Variation of WER and SER of the ASR based on the speech corpus size

Percentage from the original data set	No of speech clips	No of hours	WER%	SER%
100%	7232	4.15	12.04	21.56
80%	5743	3.30	13.84	23.96
60%	4331	2.42	14.88	25.33
50%	3612	2.02	15.05	27.50
30%	2165	1.20	17.79	32.08
20%	1435	0.86	18.34	32.30

TABLE III: Speech command classification accuracy with respect to text classification model

Text classification model	SVM	Logistic Regression	Naive Bayesian	Neural Network Embeddings
Accuracy (%)	89.7	88.9	87.2	84.5

Fig. 3 depicts the word and sentence recognition accuracy of the ASR and the accuracy of the intent classification based on the size of the training speech corpus. It is notable that even with 50% of the total corpus, a speech recognition accuracy closer to 80% could be obtained. Further, it shows that the overall accuracy of the speech command classification is following the same accuracy variation pattern compared to the ASR. After exceeding 85% word recognition and 75% sentence recognition accuracy, the intent classification reports a stable accuracy. Hence, we can suggest that the accuracy of the ASR is the major determining factor of the overall performance of the speech command classification system.

The speech command classification accuracy variation with respect to different approaches of text classification shows that SVM outperforms the other three models. It is due to the SVM's ability to perform well even with a small amount of training corpus. The Neural Network Embedding approach shows the least accuracy due to the inherent performance degradation of Neural Networks when a small training corpus is used. It is notable that the speech command classification accuracy values show only a slight variation from each other despite the text classification model used. Thus, we claim that the accuracy of the ASR is the determining factor of the overall performance of the speech command classification system.

In Fig 4, we have plotted the speech command classification accuracy against the word recognition accuracy per utterance; i.e., the fraction of correctly identified words out of the total number of words in an utterance. Once the word recognition accuracy per utterance exceeds 60%, the speech command classification accuracy increases significantly. Typically, an utterance consists of few words that have a higher weight on determining the intent. The rest of the words are commonly found in most utterances. Failing to correctly identify those significant words in an utterance has resulted to drop the speech command classification accuracy even though more than 70% of words are correctly identified. The same word is pronounced in different styles based on the context. This also can be presented as a reason for the variation of classification accuracy based on the word recognition rate per utterance.

Since our approach involves intermediary speech-to-text conversion, the WER and SER of the ASR are essential to analyze. Table 5 contains a comparison of the obtained performance results with few prior continuous Sinhala speech recognition systems. In addition, the performance comparison with some ASR systems developed for few other low-resourced Indo-Aryan languages is included in Table 6. Both cases show that our approach of using GMM-HMM based

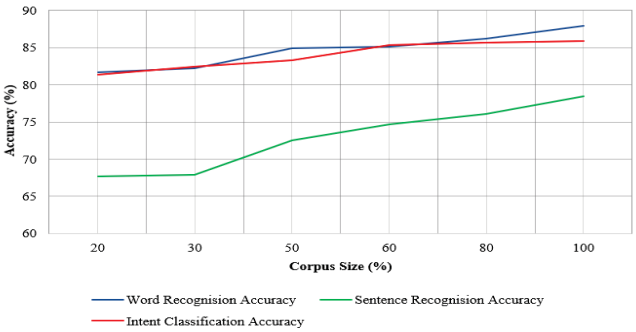

Fig. 3: Accuracy variations based on the speech corpus size

TABLE IV: Comparison with prior continuous Sinhala speech recognition systems

Speech recognision system	**ASR technique and data set**	**WER %**	**SER %**
Manamperi et al. [12]	HMM, 50 Sinhala songs, 2.63 hours, 85 speakers	-	5.7
Nadungodage et al. [13]	HMM, 983 distinct continuous Sinhala sentences, single speaker	3.86	24.26
Our system	HMM-GMM, Free-form Sinhala commands, 4.15 hours, 120 speakers	12.04	21.56

acoustic modeling in the ASR contributes to obtain a comparatively good performance in continuous speech recognition for Sinhala spoken language.

The IVR system presented by Manamperi et al. [12] shows a higher sentence recognition accuracy compared to our ASR system. The reason behind this variation can be explained as our speech corpus based on banking domain has longer utterances compared to the song recognition data set used by them. The continuous Sinhala speech recognizer by Nadungodage et al. [13] has achieved a very small WER as a result of its speech corpus collected only from a single speaker. The higher accuracy rates of the Bengali language ASR system by Zinnat et al. [16] is due to the utilization of a new feature named Local Features together with MFCCs. The ASR system presented by Kumar et al. [26] has used a training speech corpus collected only from 12 speakers. It can be presented as a reason for its lesser SER compared to our system.

VIII. Conclusion and Future Work

In this paper, we presented our domain-specific speech command classification system for Sinhala language. It detects the intent of a Sinhala utterance using ASR followed by text classification. The performance comparison of our ASR with prior ASRs developed for Sinhala language revealed that our GMM-HMM based ASR has promising results with a higher rate of accuracy. Our speech command classification system showed an accuracy of 89.7% in detecting the intent of a Sinhala utterance while outperforming the previous direct

TABLE V: Comparison with ASRs developed for other Indo-Aryan languages

Speech recognition system	Language	ASR technique	WER%	SER%
Zinnat et al. [16]	Bengali	HMM	7.50	8.50
Chowdhury et al. [25]	Bengali	HMM	28.62	-
Kumar et al. [26]	Hindi	HMM	12.99	9.07
Mohanty et al. [27]	Oriya	GMM-HMM	-	21.77
Our system	Sinhala	GMM-HMM	12.04	21.56

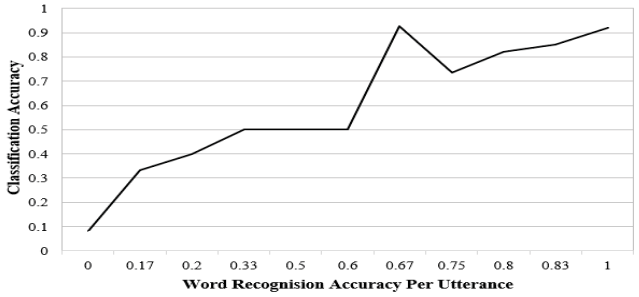

Fig. 4: Variation of speech command classification accuracy based on the word recognition accuracy per utterance

classification approach. Being the first attempt on both ASR and speech command classification for Sinhala; we suggest that our work is a positive approach and a viable solution to gain a significant improvement in intent detection of free-form Sinhala speech commands. As future work, we will extend this speech command classification system and develop a speech dialog system for the Sinhala language.

ACKNOWLEDGEMENT

The authors of this paper would acknowledge the reviewers for their valuable comments, University of Moratuwa senate research committee grant for supporting this research and all the people who participated in the data collection.

REFERENCES

[1] "Google Assistant, your own personal Google", Assistant.google.com, 2019. [Online]. Available: https://assistant.google.com. [Accessed: 19- Sep- 2019].

[2] "What is Alexa? Amazon Alexa Official Site", Developer.amazon.com, 2019. [Online]. Available: https://developer.amazon.com/alexa. [Accessed: 19- Sep- 2019].

[3] D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Jayasena, and U. Thayasivam, "Domain specific intent classification of Sinhala speech data," 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, pp. 197–202.

[4] W. S. N. Dilshani, S. Yashothara, R. T. Uthayasanker and S. Jayasena, "Linguistic Divergence of Sinhala and Tamil Languages in Machine Translation," 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 2018, pp. 13-18.

[5] Y. Karunanayake, U. Thayasivam, and S. Ranathunga, "Transfer learning based free-form speech command classification for low-resource languages," in Proc. of ACL 2019, Student Research Workshop, 2019.

[6] J. Rao, F. Ture, and J. Lin, "Multi-task learning with Neural Networks for voice query understanding on an entertainment platform," 24th ACM SIGKDD International Conference on Knowledge Discovery & Data, 2018, pp. 636–645.

[7] S. Yaman, L. Deng, D. Yu, Y. Wang and A. Acero, "An integrative and discriminative technique for spoken utterance classification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 6, pp. 1207–1214, Aug. 2008.

[8] L.R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Dec. 2011, IEEE Signal Processing Society.

[10] S. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK book, 2002

[11] M. A. Zissman, "Automatic language identification using gaussian mixture and hidden Markov models," 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, MN, USA, 1993, vol.2, pp. 399–402.

[12] W. Manamperi, D. Karunathilake, T. Madhushani, N. Galagedara, and D. Dias, "Sinhala speech recognition for interactive voice response systems accessed through mobile phones," 2018 Moratuwa Engineering Research Conference (MERCon). IEEE, 2018, pp. 241–246.

[13] T. Nadungodage and R. Weerasinghe, "Continuous sinhala speech recognizer," Conference on Human Language Technology for Development, Alexandria, Egypt, 2011, pp. 2–5.

[14] W. Amarasingha and D. Gamini, "Speaker independent sinhala speech recognition for voice dialling," Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on. IEEE, 2012, pp. 3–6.

[15] J. Nallathamby, K. Kariyawasam, H. Pullaperuma, D. Vithana, and S. Jayasena, "deBas: a Sinhala Interactive Voice Response (IVR) System," [Online]. Available: http://dl.lib.mrt.ac.lk/handle/123/8061. [Accessed: 1st July 2019].

[16] S. B. Zinnat, R. M. A. Siddique, M. I. Hossain, D. M. Abdullah and M. N. Huda, "Automatic word recognition for Bangla spoken language," 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT 2014), Ajmer, 2014, pp. 470–475.

[17] B. Karan, J. Sahoo and P. K. Sahu, "Automatic speech recognition based Odia system," 2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE), Bhubaneswar, 2015, pp. 353–356.

[18] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech Recognition system," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 493–497.

[19] T. Joachims, "Text categorization with Support Vector Machines: learning with many relevant features,"10th European Conference on Machine Learning (ECML'98), Chemnitz, Germany, 1998, pp. 137-142.

[20] D. Buddhika, R. Liyadipita, S. Nadeeshan, H. Witharana, S. Jayasena and U. Thayasivam, "Voicer: A crowd sourcing tool for speech data collection," 18th International Conference on Advances in ICT for Emerging Regions (ICTer), 2018, pp. 174–181.

[21] "ISO 15919", En.wikipedia.org, 2019. [Online]. Available: https://en.wikipedia.org/wiki/ISO_15919. [Accessed: 19- Sep- 2019].

[22] Y. Ahuja and S. K. Yadav, "Multiclass classification and Support Vector Machine," Global Journal of Computer Science and Technology Interdisciplinary, vol. 12, pp. 14 – 20, 2012.

[23] M. Y. H. Setyawan, R. M. Awangga and S. R. Efendi, "Comparison Of Multinomial Naive Bayes Algorithm And Logistic Regression For Intent Classification In Chatbot," International Conference on Applied Engineering (ICAE), 2018, pp. 1–5.

[24] L. Wu, A. Fisch, S. Chopra, K. Adams, A. Bordes, and J. Weston, "Starspace: Embed all the things!,"arXiv preprint arXiv:1709.03856, 2017.

[25] S. A. Chowdhury, "Implementation of speech recognition system for Bangla," Ph.D. dissertation, BRAC University, 2010.

[26] K. Kumar, R. Aggarwal, and A. Jain, "A Hindi speech recognition system for connected words using HTK," Int. Journal of Computational Systems Engineering, vol. 1, pp. 25 – 32, 2012.

[27] S. Mohanty and B. K .Swain, "Continuous Oriya digit recognition using Bakis Model of HMM," International Journal of Computer Information Systems, vol. 2, no. 1, 2011.