Domain Specific Intent Classification of Sinhala Speech Data

Darshana Buddhika, Ranula Liyadipita, Sudeepa Nadeeshan, Hasini Witharana, Sanath Jayasena
and Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
ranula.14@cse.mrt.ac.lk

***Abstract*—Building an open domain automatic speech recognition(ASR) system can be accomplished by converting voice into text and performing a text classification on top of the converted text. However, with the inherent challenges in the approach mentioned above, it is not the most feasible way of the deriving intent of speech queries in a specific domain. This paper proposes a domain-specific intent classification for Sinhala language utilizing a feed-forward neural network with backpropagation. For the purposes of this research, a Neural network is trained from Mel Frequency Cepstral Coefficients (MFCC) which are extracted from a Sinhala speech corpus of 10 hours and the performance of the system is evaluated using the recognition accuracy of the speech queries. Further, the proposed solution in the paper introduces the first-of-its-kind for domain-specific intent classification for Sinhala language.**

***Keywords*-Sinhala Speech Recognition; Feed Forward Neural Network; Speech recognition;**

I. Introduction

Speech recognition[1] has been evolved tremendously with the advancement of technology. As a result of this, speech-driven-systems are filling the gap between the human and the machine. Understanding the intent out of uttered words is an advancement of normal Automatic Speech Recognition(ASR) systems. Intent classification is used in many commercial products, such as Siri[2], Alexa[3] and Google Home[4]. This paper describes the development of a domain-specific Neural Network based intent classification system for Sinhala language.

Intent classification can be done either by converting speech to text and doing a text classification[5] or directly using a classification algorithm for extracted features of speech. The first method is widely used in many generic ASR systems, whereas the second method is more applicable for specific domains. For a specific domain, the speech vocabulary is limited and well defined.

Sinhala is an Indo Aryan language which belongs to the low resource language category. 75% of Sri Lankans use Sinhala as their mother language[6]. There are prior researches which had conducted focusing Sinhala speech recognition such as [7] [8], [9], [10], [11], [12], [13], [14] and [15]. Research [7] - [14] have used Hidden Markov Model(HMM) for the development of Sinhala ASR systems where as [15] has used Dynamic Time Warping(DTW). At the time of writing no prior research has conducted for intent classification using only extracted speech features for Sinhala language.

This paper presents a domain-specific, intent classification system for the Sinhala language based on a neural network. We have selected banking domain to conduct the experiment. In respect of banking domain, we have selected six different queries which are often used by customers in the event of interacting with the bank.

1) Requesting account balance.
2) Requesting to deposit money to a bank account.
3) Requesting to withdraw money from a bank account.
4) Bill payments.
5) Transfer of money from one account to another.
6) Credit card payments.

These queries will be addressed as “intents” in the following sections. We have recognized different ways in which each one of these intents are communicated through speech and these variations are referred to as “inflections” of a intent from this point onwards.

e.g.
Domain - Bank
Intent - Requesting Balance
Inflections - I want to know the balance, What is the balance of my account?

In the research, we have extracted MFCC features in the process of audio feature extraction because MFCC is known for its efficiency and high accuracy with less complexity[16]. Then extracted features are used to train a neural network based system. As of the time of writing, we have gathered 10 hours worth Sinhala speech data collected from 152 males and 63 females for the banking domain by using a web/smartphone based voice collection tool developed for this research. System has reached an accuracy of 74% in the experiments with the aforementioned dataset. Our paper further evaluates the training and test accuracy of intent classification against neural network, support vector machine and decision tree based systems.

The rest of the paper is organized as follows. In section II, we provide related work that has been done for Sinhala speech recognition, intent classification for other languages and speech feature extraction. Section III presents the methodology of the proposed neural network based intent classifier and the experimental method used in this research is described in section IV. In section V, we present the results of the experiment carried. Section VI contains the discussion of the results. We conclude the research paper by giving the conclusion and future work in the final section.

II. RELATED WORK

Speech recognition has gained huge popularity in the last decade. There are number of researches that had taken place to build ASR systems for different languages. Following sections evaluate them on different perspectives related to the intent classification system that we propose.

A. Speech Recognition Techniques

Nowadays, application with speech recognition is used widely and many researchers follow a variety of approaches in doing so. That includes DTW, HMM and Neural Network(NN) based approaches.

DTW- Dynamic time warping is an algorithm used for measuring the correlation between two sequences. This algorithm can spot similarities between two speech signals irrespective of their speeds and the 'similarity cost' produced by this algorithm represents the level of correlation between the two speech signals[17].

HMM - HMM approach provides statistical models that output a sequence of symbols or quantities. Each state in HMM contains a statistical distribution which provides the likelihood for a particular observed vector. To build a HMM for a word/phoneme sequence one has to use individually trained HMM for separate words/phonemes in succession. A large number of speech recognition systems tend to be based on this particular method[18].

ANN - Combination of HMMs and Artificial Neural networks (ANN) are used for acoustic modeling, language modeling as well as noise reduction which are core elements in speech recognition.[19]. Instead of combining Neural Networks with HMMs, the newest trend is to train Neural Networks 'end-to-end' for speech recognition[20]. A major player in ASR industry, Google Home uses a similar approach and they have included deep neural networks in their configuration as they possess a very large amount of speech data[4].

B. Intent Classification using Neural Networks

Majority of the researches conducted for speech recognition in the past had used the HMM and DTW approach. But modern approaches are more focused towards neural networks. Reference [21] suggests a Neuro-fuzzy system to recognize the human voice and it has a high efficiency and accuracy because it combines strong points from Neural Network and Fuzzy Logic. The system in [21] focuses on identifying 8 intents in Thai language and for each of these intents a separate feedforward neural network containing two hidden layers are used and Fuzzy logic is implemented above the outputs of networks. The system we propose includes a single neural network unlike [21].

Speech recognition system presented in [5] directly transcribes audio data to text, without requiring an intermediate phonetic representation. The system combines the deep bidirectional Long Short-Term Memory(LSTM) neural network architecture and the Connectionist Temporal Classification objective function. This approach results in a text transcript of a speech query rather than intent itself.

C. Sinhala Speech Recognition Systems

There are some research conducted for the Sinhala language based speech recognition. Systems in [7] - [14] use HMM to recognize Sinhala voice clips. A speaker independent Sinhala speech recognition system for interactive voice response systems accessed through mobile phones is presented in [7]. The research [8] is developed using the Hidden Markov Model Toolkit(HTK) to identify Sinhala digits. A combination of both ASR and text to speech is presented in a system called "debas"[9]. The paper [10] presents another HTK based speaker independent Sinhala speech recognition system for voice dialing. A continuous speaker dependant Sinhala speech recognition system is presented in [11] which is developed using HTK. The work in [11] is extended using a speaker adaptation model in [12]. The work [13] and [14] present two Sinhala speech recognition systems developed using HTK. The research[14] is trained using 10 speakers. Reference [15] shows a Sinhala ASR system developed using dynamic time warping(DTW). This system has the capability of identifying frequently used words.

D. Speech Feature Extraction

A compact, paramedic representation of the input speech wave (speech signal) can be generated using signal processing. This process is called feature extraction. This feature extraction process reduces irrelevant data from the signal and captures its essential elements. These parameters are then used for further analysis. Each speech signal has been implanted some elements which are unique to an individual speaker. Various feature extraction techniques can be used to remove these elements from the signal. These represented parameters are used in speech recognition, speech synthesis, speaker recognition, speech coding, voice modification, and enhancement. These extracted features can be benchmarked using a criterion as follows.

- Easiness to generate
- Stable over time
- Not open to mimicry
- Changing little from one speaking environment to another.

Techniques like MFCC, Relative Spectral Filtering(RASTA), Perceptual Linear Predictive(PLP), Linear Predictive Coding(LPC) and RASTA-PLP are widely used approaches for extracting features from the speech.

1) LPC: A powerful technique used in speech recognition. The robustness and accuracy of the technique have made it popular in the industry. In LPC observational vectors are generated using a frame-based inspection of the speech wave. LPC is also used for signal compression and synthesis[16].

2) RASTA: This feature extraction methodology was first developed aiming to subside the extra noise in the speech signals. It intensifies the quality of speech with background noise as well as reduces the impact of the noise[16].

3) PLP: PLP follows the same approach as LPC. It uses a short-term spectrum of speech and performs several psychophysically based transformations on the short-term spectrum of speech and improves it [2]. It applies spectral analysis on speech signal vector with frames of N samples with Nband filters[22].

4) MFCC: MFCC is considered as the state of the art in feature extraction. Human ear's critical bandwidth along with frequency is the basis for this particular feature and it tries to imitate the human auditory system. After performing feature extraction on the input signal, its frequency information is converted to a set of coefficients and the process of conversion is known for its fast and reliable computation methodology[16].

III. Methodology

In this research, we are using a multilayer feed-forward neural network[15] for classifying and recognition of Sinhala speech utterances. The whole process of recognizing Sinhala audio signal consists of two parts as described in subsections below.

A. Feature Extraction

We are using MFCCs for feature extraction because MFCC is very efficient and accurate with low complexity[16]. The following steps are used to calculate MFCC features.

- Framing the signal into short frames.
- Calculating periodogram estimate of the power spectrum for each frame.
- Applying the Mel filter-bank to the power spectra sum the energy in each filter.
- Taking the logarithm of all filter-bank energies.
- Taking the Discrete Cosine Transform of the log filter-bank energies.
- Keeping Discrete Cosine Transform coefficients 13, discard the rest.

The high level process of MFCC feature extraction can be identified in Fig. 1.

B. Multilayer Feed-forward Neural Network

Extracted MFCC features are used as inputs after preprocessing. Original feature metrix derived from MFCC have a shape of 13*N, where N depends on the length of the audio signal. Then these features are reshaped to fit into the neural network.

Feed-forward neural network with multiple layers has used as the neural network model and backpropagation is used as the learning algorithm. Fig. 2 shows the high-level structure of the neural network model with the number of nodes in each layer. First, the input layer takes the MFCC features as input vectors and forward those features to the next layer with different weights.

The process inside a node in a hidden layer and the output layer is shown in Fig. 3. The output (y) from each node calculate the summation of data (x), weights(w) and the bias (b). Equation (1) shows the output function.

$$y = \Sigma x_i w_i + b \tag{1}$$

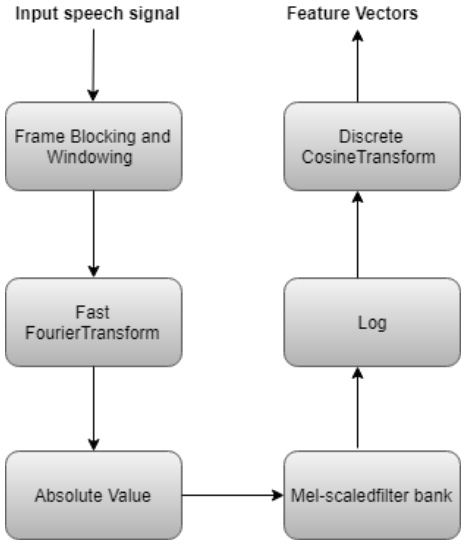

Figure 1. MFCC Feature extraction process can be broken down into six major steps and it converts the given input speech signal into a sequence of feature vectors.

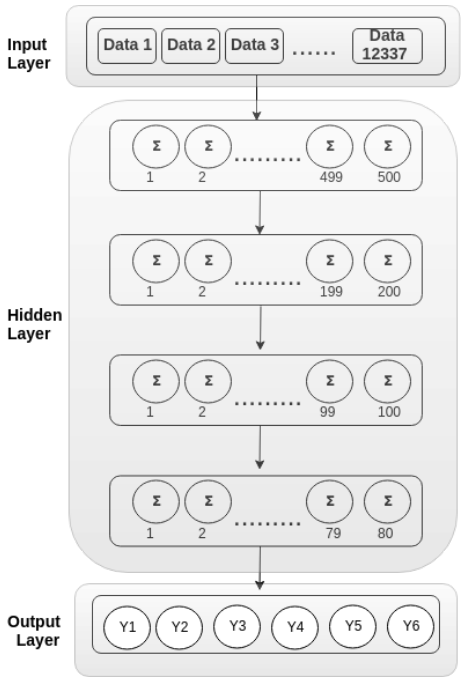

Figure 2. Input layer has 12,337 nodes since that is the size of the MFCC feature vector fed to the network. Output layer has 6 nodes representing the intents.

Each node in the hidden layers uses relu (rectified linear unit) as the activation function. This function takes an input(y) and returns the same value if the value is greater than zero. Otherwise, it returns zero. According to (2) if the input is y,

$$y^i = max(y, 0) \tag{2}$$

In the last layer, we use six neurons with sigmoid activation to get the probability of being in each class. This model uses categorical cross-entropy to calculate the error between the predicted result and the actual result.

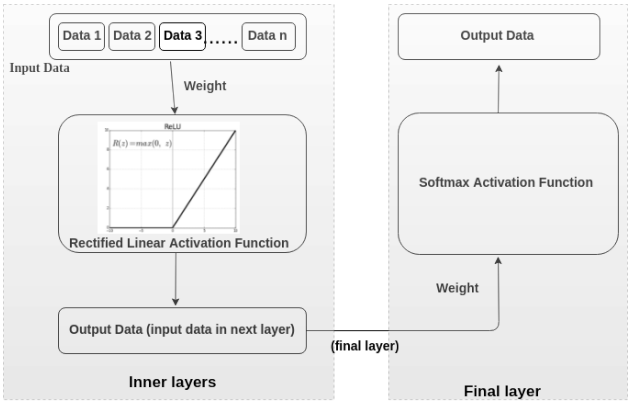

Figure 3. Structure of each node in hidden layer and output layer is shown in the graph. Nodes in the hidden layers uses relu(rectified linear unit) as the activation function and output layer nodes uses Softmax Activation Function

Table I
ACCURACY AND LOSS OF THE TRAINING SET AND VALIDATION SET

	Sample	Accuracy%	Loss%
Training	6396	94.14	0.1736
Testing	1600	74.37	0.9416

IV. EXPERIMENTAL METHOD

We did a comparison of intent classification accuracy acquired by neural networks with the accuracy of decision tree approach and support vector machine approach. For the experiment, we have utilized a speech corpus of 7996 Sinhala speech samples which worth 10 hours, for 6 basic intents in banking domain as mentioned in the section I. Each of the intents consists of different inflections of that intent and for the 6 intents there were 38 different inflections altogether. Model is expected to classify any inflection to its basic intent. Speech data was divided into two sets for training and testing purposes with 80%, 20% ratio respectively. Then the accuracy of intent classification was compared among the three models.

V. RESULTS

The error rate of the neural network is based on the number of misclassified speech data. The percentage loss and accuracy for training and testing are shown in Table I. To calculate loss, categorical cross-entropy is used. For better performance loss should be minimized.

Fig. 4 represent the categorical cross-entropy of the system based on epoch values. The number of epoch value for maximum performance was decided by examining the minimum validation loss epoch value. According to Fig. 4, we chose 60 as epoch value because we got best validation performance at epoch 60.

Fig. 5 and Fig. 6 show the classification accuracy of training and test samples. Six selected commands are represented as 0-5. Confusion matrix plot the classified number of speech data against the actual prediction vs expected classification. According to the Fig. 7, training data sample has more accuracy compared to test data sample.

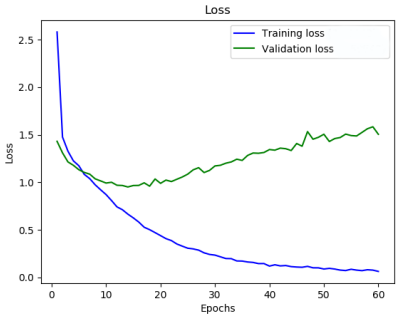

Figure 4. Training loss and validation loss of the model for 60 epochs

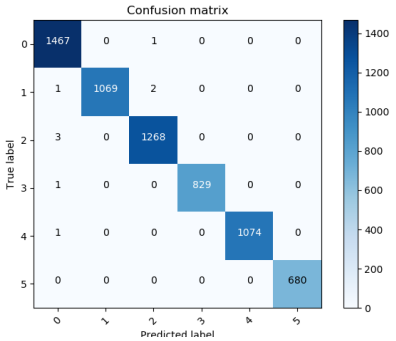

Figure 5. Confusion matrix for training data shows that only a few samples has classification issues.

Fig. 7 shows the performance of the intent classification system proposed by this research. According to Fig. 8 training accuracy is 92% and the test accuracy is 74%.

The intent classification accuracy of Neural Network(NN), Support Vector Machine(SVM) and Decision Tree(DT) approaches are shown in Table II. According to Table II, Neural network approach has shown better performance than support vector machine and decision tree approach in domain-specific intent classification for Sinhala language.

VI. DISCUSSION

To identify the intent, a majority of the existing technologies in ASRs are converting the voice sample into a text and then do a text classification. While it is arguably

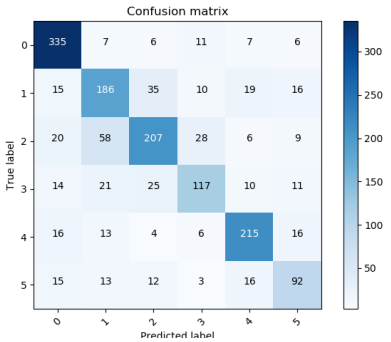

Figure 6. Confusion matrix for test data shows majority of the samples are classified properly

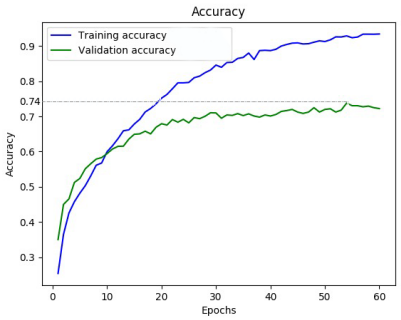

Figure 7. Train and test accuracies of the model are show in the graph and it shows the highest of 74% for the test data set.

Table II
COMPARISON OF NN, SVM AND DT METHODS

	NN	SVM	DT
Accuracy	74 %	52 %	64 %

the best approach in the open domain speech recognition, it is an overwhelming approach for a less complicated domain-specific speech recognition system due to several reasons. To convert a speech into a text, one has to build the language model and acoustic model or find a data set large enough and feed into a deep neural network to figure those models on its own. That being said, finding/building a data set that large is, in fact, an infeasible task for a low resourced language. Building an acoustic model and language model requires the combination of both technical knowledge and language expertise. Even if a reasonable corpus is collected and required models are built, to combine them and build the system one has to understand and configure a framework like kaldi[23].

With the approach we propose complexity aligned with ASR is narrowed down to a less complicated intent classification problem. However, the system has its own limitations and cannot be identified as an alternative to open domain ASR solutions. This system performs well in specific domains where required intents are defined clearly. We have tested the system with 6 different intents as aforementioned. In the data set each intent contained 4-6 inflections. Since the number of intents equals to classes in the classification, having a higher number of intents for classification reduces the accuracy of the system. However, the system handled a higher number of inflections(38) under a limited number of intents(6) with a significant accuracy in the experiments. During the model development for intent classification, we have identified a few trends which will help further researches conducted in this area.

A. Number of data samples

Fig. 8 shows the accuracy of intent recognition increases as the number of speech utterances increases. As shown in Fig. 8, it is possible to reach 74% of accuracy with a 10-hour speech data set.

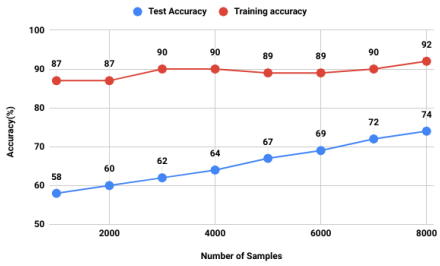

Figure 8. Model accuracy increases with the number of data samples.

Table III
ERROR DISTRIBUTION OF THE RECORDED UTTERANCES

	Over Recorded	Stopped halfway	High noise profile
Number of Clips	147	81	238

B. Data Preprocessing

In the earlier steps of model development, we tried training the neural network with raw recordings in the corpus. Even if we tried several approaches the accuracy of the model did not exceed 66%. Then we did a preprocessing of the corpus. According to Table III, there were 147 over recorded samples in the corpus. This has occurred as a result of users pausing for a very long time either in the beginning, middle or at the end of the recording which finally led the sample to be over-recorded. There were some speech recordings that did not contain the full inflection, users have stopped halfway of the recording. In the speech corpus, we found 81 such recordings. As the process of data collection was not conducted in a noise-free environment, we had to remove 238 speech samples due to the impact of heavy noise. Finally, we had 7994 speech recordings that had passed all the quality requirements.

After performing this data preprocessing step, our model accuracy improved to 73% and it shows the importance of having preprocessed corpus for the system we propose.

C. Data Imbalance of the Classes

The number of samples for different intents are not equal in the dataset and as the table IV shows intent 1 and intent 6 has a difference of 1000 voice samples. This critically affects the performance of the system. As the number of inflections varies from intent to intent this is a common characteristic of the corpus. Hence to reduce this it is important to have a similar number of inflections per an intent. To overcome the imbalance we calculated a weight for each class and trained the module according to the weights and it has improved the accuracy from 73% to 74%.

VII. CONCLUSION AND FUTURE WORK

In this paper, we propose a feed-forward neural network approach which utilizes MFCC features for domain-

Table IV
CLASS WEIGHT DISTRIBUTION

Intent	1	2	3	4	5	6
Number of voice Clips	1839	1340	1602	1028	1355	830
Class weights	0.72	0.98	0.83	1.29	0.99	1.61

specific intent classification in Sinhala language. A banking domain related data set is used to train the system and the accuracy of the aforementioned neural network(NN) approach was 74%. We identified that NN is the most suitable approach compared to support vector machine and decision tree approaches in this context. We are planning to identify the impact of different feature extraction techniques like Linear predictive coding(LPC), Linear Prediction Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLP), Relative Spectral Filtering (RASTA) in intent classification for Sinhala language as future work.

ACKNOWLEDGMENT

Authors of this paper acknowledge University of Moratuwa senate research committee grant for supporting this research.

REFERENCES

[1] K. Ahammad and M. M. Rahman, "Connected bangla speech recognition using artificial neural network," *International Journal of Computer Applications*, vol. 149, no. 9, 2016.

[2] S. Team, "Deep learning for siri's voice: On-device deep mixture density networks for hybrid unit selection synthesis," *Apple Machine Learning Journal*, vol. 1, 2017.

[3] "How amazon alexa works," Available at https://channels.theinnovationenterprise.com/articles/how-amazon-alexa-works (2018/07/19).

[4] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin *et al.*, "Acoustic modeling for google home," *INTERSPEECH-2017*, pp. 399–403, 2017.

[5] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[6] "Sri lanka population 2018," Available at http://worldpopulationreview.com/countries/sri-lanka-population/ (2018/07/19).

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

[8] "kathana sinhala speech recognition system," Available at http://archive.is/BGvEp (2018/07/19).

[9] "debas: a sinhala interactive voice response (ivr) system," Available at http://dl.lib.mrt.ac.lk/handle/123/8061 (2017/07/18).

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

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

[12] T. Nadungodage, R. Weerasinghe, and M. Niranjan, "Speaker adaptation applied to sinhala speech recognition." *Int. J. Comput. Linguistics Appl.*, vol. 6, no. 1, pp. 117–129, 2015.

[13] ——, "Efficient use of training data for sinhala speech recognition using active learning," in *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*. IEEE, 2013, pp. 149–153.

[14] W. Samankula and N. Dias, "Designing an automatic speech recognition system to recognize frequently used sentences in sinhala," 2013.

[15] P. Priyadarshani, N. Dias, and A. Punchihewa, "Dynamic time warping based speech recognition for isolated sinhala words," in *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on*. IEEE, 2012, pp. 892–895.

[16] K. Gupta and D. Gupta, "An analysis on lpc, rasta and mfcc techniques in automatic speech recognition system," in *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*. IEEE, 2016, pp. 493–497.

[17] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[18] J. A. Bilmes, "What hmms can do," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 3, pp. 869–891, 2006.

[19] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer Science & Business Media, 2012, vol. 247.

[20] K. Kawakami, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, PhD thesis. Ph. D. thesis, Technical University of Munich, 2008.

[21] K. Srijiranon and N. Eiamkanitchat, "Thai speech recognition using neuro-fuzzy system," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference on*. IEEE, 2015, pp. 1–6.

[22] U. Shrawankar and V. M. Thakare, "Techniques for feature extraction in speech recognition system: A comparative study," *arXiv preprint arXiv:1305.1145*, 2013.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.