Sādhanā (2022)47:284
https://doi.org/10.1007/s12046-022-02058-z

A study on conventional and syllable-based approaches for automatic speech recognition in Malayalam

JASMIN S[1], ASHISH ABRAHAM SAMUEL[2] and RAJEEV RAJAN[2,*]

[1]L &T Technology Services, Mysore, Karnataka, India
[2]College of Engineering, Trivandrum, APJ Abdul Kalam Technological University, Thiruvananthapuram, India
e-mail: rajeev@cet.ac.in

MS received 15 June 2022; revised 19 October 2022; accepted 4 November 2022

Abstract. This paper investigates the conventional and syllable-based ASR systems for a low resource south Indian language, Malayalam. The standard Kaldi frame work is employed. While the first approach uses word-phoneme lexicon, the second approach uses syllable-phoneme lexicon as pronunciation dictionary. Mel-frequency cepstral coefficient features of the audio corpus are extracted, and acoustic modeling is done using the Gaussian mixture model—hidden Markov model and deep neural network (DNN). The systems' performance dependence on different factors like the type of modeling and alignment algorithms employed are studied. The number of hidden layers and units are varied, and the result is analysed. The fine-tuning of phoneme positions plays an integral part in the Kaldi speech recognition toolkit recognition process. The syllable-based study is conducted using a novel phonetic analyser, Mlphone. The analysis shows that Kaldi performed well for phoneme-level DNN acoustic modelling, providing a lower word error rate of 2.86% than the syllable-based model.

Keywords. Speech recognition; syllable; deep neural network; word error rate.

1. Introduction

Automatic speech recognition (ASR) is a present-day technology that enables humans to communicate with machines directly through speech. ASR systems are used in many applications since communication with the devices through speech is natural and user-friendly. Depending on the variations in speech dialects, the complexity of the language, and the environmental conditions, the effectiveness of speech recognition varies, and the development of an ASR system becomes more challenging. The ASR systems developed using the publicly available software tool Kaldi possess faster real-time recognition, and a high-quality network [1]. The speech recognition in Kaldi is done mainly by predicting the position of the phonemes from the audio information and the provided transcription. Hence, the speech recognition system's accuracy depends on the exactness of the phoneme positions. In the proposed study, two approaches, namely, word-phoneme and syllable-phoneme lexicons, are attempted using Kaldi framework. The transcription accuracy is investigated using a standard Malayalam speech corpus. Modelling similar polyphones with the same HMM through state clustering is done using a phonetic decision tree. The leaves of the decision tree give the desired state clusters. In this work, the decision trees obtained from the GMM-HMM are used to train a hybrid Deep Neural Network - hidden Markov model (DNN-HMM).

Malayalam is a low resource language for which only a few studies [1–5] have been performed so far. The survey in [6] shows the effect of CNN, DNN and RNN on ASR systems and proves that the DNN-HMM models provide better recognition accuracy than traditional GMM-HMM models. The low accuracy of the existing speech recognition systems and the possibility of exploring DNN-HMM models motivated us to analyse recognition accuracy using the mentioned approaches.

2. Automatic speech recognition models

using Kaldi, a DNN-HMM hybrid system is implemented for continuous Malayalam speech. Posterior probabilities of each context-dependent state are predicted using DNN with acoustic cues. The processes involved in developing a DNN-HMM model consist of feature extraction, training of monophone model, training of triphone model with delta features and delta-delta features, training a triphone model with linear discriminative analysis (LDA) and maximum likelihood linear transform (MLLT), speaker-adapted training (SAT) and training the final DNN-HMM model. MFCCs are computed with a 25 ms window shifted by 10 ms. Each phoneme is modelled with five states HMM.

*For correspondence
Published online: 20 December 2022

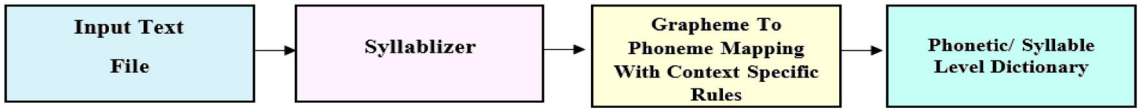

Figure 1. Block diagram of *mlphon*-based transcription approach.

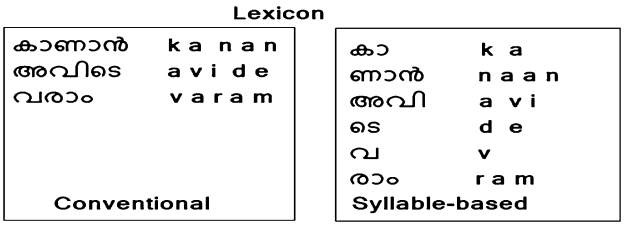

Figure 2. Lexicon for Conventional and Syllable-based ASR.

Language models are developed using standard bigram models, which aim to refine the thorough recognition of informal speech. Statistical language models assign probabilities to the sequence of words. The result of language modelling is an *ARPA* file containing all the words provided in the dictionary file and their recognition probabilities. The proposed work develops two different language models for two dictionaries. Conventional transcription follows the standard ASR procedure with a word-phoneme lexicon.

In the second syllable-based approach, a syllable-phoneme lexicon as a pronunciation dictionary is employed. It uses a phonetic analyzer tool, *mlphon* [5] for syllabification. *mlphon* gives a rule-based syllable level lexicon for the Malayalam language, based on which the transcriptions are written, and a speech recognition system is developed using Kaldi. A sample of lexicon is shown in figure 2.

Mlphon is a novel Malayalam phonetic analyzer that provides the grapheme sequence's phonetic characteristics. One of the primary use of the phonetic analyzer is to create a phonetic lexicon that can be used in automatic speech recognizers and other text-to-speech converters. The tool is implemented based on finite-state transducers (FST), consisting of a finite number of states linked by transitions labelled with an input/output pair. A block diagram of the *mlphon* phonetic analyzer is shown in the figure 1. We followed the implementation given in [5]. The g2p implementation uses Stuttgart finite state transducer (SFST) and Helsinki finite-state technology (HFST). The g2p mapping of Malayalam in *mlphon* is implemented by considering the rules to handle *schwa* (a neutral vowel) addition at beginning/end/middle of words depending on the presence of *chillus* and *virama* [5], phonetic variations due to the context of a certain sequence of consonants, contextual nasalization [5]. The g2p conversion problem is typically broken down into three subproblems, namely sequence alignment, model training and decoding. While the sequence alignment aligns the grapheme and phoneme sequence pairs in a training dictionary, model training generates a model to create new pronunciations for novel words.

State transitions are labelled with input and output symbols in a finite-state transducer. Therefore, a path through the transducer encodes a mapping from an input symbol sequence to an output symbol sequence [7]. An FST is defined8 as 6-tuple (Σ, Γ, Q, q_0, F, δ),

where: Σ = finite, non empty set of input symbols
Γ = finite, non empty set of output symbols
Q = finite, non empty set of states
q_0 = initial state, $q_0 \in Q$
F = set of final states, F $\subseteq$ Q
δ = transition function, which can be represented with

$$\delta \subseteq Q \times (\Sigma \cup \epsilon) \times (\Gamma \cup \epsilon) \times Q. \quad (1)$$

An FST representing a simple pronunciation mapping that accepts two words is shown in figure 3. States are represented as circles and marked with their unique number. The initial state is represented by a bold circle, and the final state by double circles. An input symbol i, and an output symbol, o, are marked on the corresponding directed arc as i :o. ϵ is a special symbol that indicates the generation of an output corresponding to an empty input string [5].

A syllablizer consists of a word filter and a syllable. Word filter is the first level transducer in *mlphon*, which accepts the text input file and pre-processes it by adding word wrapping tags. Syllable FST helps distinguish the syllables' beginning and end with tags. For the words split at the syllable levels, g2p mapping is done. Hence, the syllablizer acts as a pre-processor for g2p mapping. Ml2ipa is the FST that converts the Malayalam text to IPA mode. It makes use of g2p FST and a tag filter for this purpose. In this work, a lexicon is created in the form of syllable-phonemes after syllabification of words in the dictionary

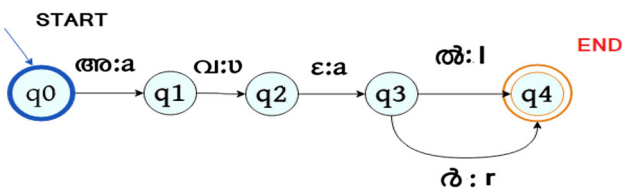

Figure 3. FST representing a simple pronunciation mapping that accepts two words.5

Table 1. Dataset description of 4 and 5 h of Corpus.

	4 h		5 h	
Data	Speakers	Data	Speakers	Data
Train	24	3 h	38	4 h
Test	9	1 h	9	1 h

using *mlphon*. The obtained lexicon consists of 1072 and 1676 syllables for the 4 hrs and 5 hrs corpus, respectively.

3. Performance evaluation

3.1 *Database*

We used Indic TTS-IITM [9] data corpus for the proposed study. The corpus was collected by TTS Consortium led by IIT Madras. Database and text-to-speech synthesizers are built for all the 13 languages. We have conducted the experiment with Malayalam subset corpus. Malayalam corpus consists of a compilation of Malayalam sentences spoken by male and female speakers. The database description is given in table 1. Speaker utterances are typically proverbs from Malayalam, newspaper reports, short stories and blogs collected from various resources. Audio files from 19 female and 14 male speakers are used in the training set, and five female and four male speakers are used for the test set during for the four hour corpus. The overall model is speaker-independent.

The details related to the Indic corpus can be referred to in [9]. The recording is carried out in noise or echo-free environment by professional speakers(male and female) to maintain a constant pitch and prevent stress. Further, to avoid the fatigue of the speaker, a break is given every 45 minutes. Later the recorded sentences are split at the sentence level. Also, measures have been taken to maintain the same conditions and voice characteristics across the multiple recording sessions. The recorded speech files are stored in .wav format in the database with a sampling rate of 48 KHz.

3.2 *Experimental setup*

The experiment is carried out in a system with dual core Intel Core i3 processor and Ubuntu 18.04 (64-bit operating system). The results are calculated using word error rate (WER). The bi-gram language models are trained on the Malayalam text corpus of online newspapers. The SRI Language Modeling Toolkit (SRILM) is used to build the language model. Then, by applying a set of g2p rules, phonetic transcriptions for the words from the vocabulary were rendered automatically. At first, the experiment is conducted using acoustic models from GMM-HMM. Later, acoustic models based on DNN are generated. Our acoustic model DNN will have input nodes corresponding to the dimensions of our audio features and output nodes corresponding to senome labels. The input features to the neural network are MFCC (spliced) + LDA + MLLT + fMLLR, 40-dimensional features. We initialize the neural net with a single hidden layer; we increased the number of hidden layers later in training. Three hidden layers with 50-200 units in each are part of the proposed DNN architecture. The details of are given in table 2.

Table 2. DNN implementation.

Tool/parameter	Value
DNN codebase	nnet2
Hidden layers	3
Activation function	p_{norm}
num_{epocs}	8
Mini-batch size	512
Learning rate	0.003

A monophonic model is trained initially. Later, the experiment is extended to triphone moel by considering the meaning of left and right phonemes. Triphone models are more powerful than monophonic models because speech varies depending on the context. In addition, two parameters are transferred during the training of the triphone model: no leaves in the decision tree and the total number of Gaussians in the model across all states. As we refine the models, the values are increased until optimal results are obtained.

3.3 *Results analysis*

Transcrption examples are given in figures 4 and 5 for conventional and syllable-based approaches, respectively. The experimental results of various phases are tabulated in table 3. By examining the fourth column of the table (Conventional, four hrs of data), it can be noted that the results are improved with a WER of 2.95% for hybrid DNN than 4.56% for the baseline monophone model with number of Gaussians N_f=500. Even in the triphone model, the improvement in WER can be noticed from the table for

```
mlf_01640_00015090452    ന അടിയിൽ
mlf_01640_00124672388    ഡാറ്റ കണക്ഷൻ നഷ്ടമായി
mlf_01640_00170003835    നിങ്ങളുടെ ലക്ഷ്യസ്ഥാനം വലതുവശത്താണ്.
mlf_01640_00170219893    രണ്ട് കിലോമീറ്ററിൽ
mlf_01640_00185267378    നേരേ ഇടത്
mlf_01640_00202553470    ഒരു യു-ടേൺ എടുക്കുക
```

Figure 4. Conventional transcription.

```
mlf_01640_00015090452   ന അ ടി യിൽ
mlf_01640_00124672388   ഡാ റ്റ ക ണ ക്ഷൻ ന ഷ്ട മാ യി
mlf_01640_00170003835   നി ങ്ങ ളു ടെ ല ക്ഷ്യ സ്ഥാ നം വ ല തു വ ശ ത്താ ണ്
mlf_01640_00170219893   ര ണ്ട് കി ലോ മീ റ്റ റിൽ
mlf_01640_00185267378   നേ രേ ഇ ട ത്
mlf_01640_00202553470   ഒ രു യു-  ടേ ൺ എ ടു ക്ക ക
```

Figure 5. Syllable-based transcription.

Table 4. Best schemes with respect to the WER results

Approach	4 h	5 h
Conventional ASR	2.86	2.76
Syllable-based	14.14	13.86

the customized model parameters. A similar trend is observed for N_f=1500 and 5000. It is observed that there is a significant improvement in the system's performance for refined models with different training algorithms for HMM model, as expected. The phoneme-to-audio alignments obtained from the HMM model with the best WER are used to train the DNN model. The performance of the DNN model depends significantly on the HMM model and the level of tuning. DNN performance was found to rely primarily on two factors: The effectiveness of the previously trained HMM-GMM and neural net tuning. GMM-HMM often results in poor alignment, decreasing the overall model's efficiency regardless of the tuning used. The output variance can also be observed from table 3 with different data sizes; 4hrs and 5 hrs of data.

In the syllable-based approach, it is noted that WER is high as compared to the conventional method. This can be examined by comparing columns 4 with 6 and 5 with 7 of table 3. As corpus size increases, WER decreases, as seen in the conventional approach. For the syllable model, the number of words in the dictionary is small but repeated several times in the corpus. While for the word level corpus, the repetition of the words is significantly less, the number of words in the dictionary is very high. For the word level model, the number of words in the dictionary is 27752, while 7546 for the syllable level model is for 4 hrs of the corpus. Hence, it is expected that the syllable model will perform better than the phoneme model with increased corpus size. So, the experiment is repeated by increasing the corpus size, and the results are listed in table 3. With the increase in the corpus, the WER decreased for the syllable model, but the drift is marginal compared to the conventional ASR. But as model moves from baseline to hybrid significant change in WER is observed, For example, by examining the 6th column of table 3, we can observe that WER is reduced from 39.96% to 19.31% for N_f=500. The best results from the two approaches are re-reported in table 4. The conventional approach reports a WER of 2.87 with a significant margin over the syllable-based approach.

Table 3. Results of conventional and syllable-based ASR for 4 hrs and 5 hrs of Malayalam Corpus. N_l and N_g represent number of leaves and number of Gaussians respectively. WER is given in percentage.

			Conventional		Syllable	
Model	N_l	N_g	WER_4	WER_5	WER_4	WER_5
Monophone		500	4.56	4.09	39.96	36.68
Tri1 (delta)	500	2750	3.46	3.20	25.67	24.93
Tri2a ($delta + deltadelta$)	500	2750	3.60	3.11	25.54	24.01
Tri2b ($LDA + MLLT$)	500	2750	3.75	3.14	22.93	21.45
Tri3b ($LDA + MLLT + SAT$)	500	2750	3.94	3.23	23.26	22.53
Hybrid (DNN)	500	2750	**2.95**	**3.02**	**19.31**	**19.12**
Monophone		1500	4.04	3.76	35.72	33.26
Tri1 (delta)	1500	15000	3.95	3.29	20.40	19.86
Tri2a ($delta + deltadelta$)	1500	15000	3.80	3.23	19.78	19.12
Tri2b ($LDA + MLLT$)	1500	15000	3.92	3.33	17.6	16.40
Tri3b ($LDA + MLLT + SAT$)	1500	15000	4.44	3.46	14.14	13.86
Hybrid (DNN)	1500	15000	**2.86**	**2.76**	**18.98**	**16.82**
Monophone		2000	3.94	3.80	36.42	34.44
Tri1 (delta)	5000	20000	4.44	4.01	20.44	21.32
Tri2a ($delta + deltadelta$)	5000	20000	4.38	4.43	20.57	20.11
Tri2b ($LDA + MLLT$)	5000	20000	4.91	4.34	17.36	17.59
Tri3b ($LDA + MLLT + SAT$)	5000	20000	5.60	4.93	13.81	13.89
Hybrid (DNN)	5000	20000	**2.87**	**2.79**	**19.01**	**18.17**

WER for hybrid model is highlighted in bold

Table 5. Comparison of WER from previously reported works on Malayalam ASR.

ASR model	WER (%)
Deekshita et al. [10]	34.20
Lavanya et al. [1]	34.40
Lekshmi et al. [11]	10.00
Kavya et al. [5]	9.60
Proposed approach (Syllable)	13.86
Proposed approach (Conventional)	2.76

Although there have been previously published works on ASR for continuous Malayalam speech,1,10,11 each one was tested using private datasets described in respective papers. The lexicon size for each work is different. Nevertheless we present a comparison of previously reported WERs with ours in table 5. Our analysis shows that conventional approach outperforms syllable-based for small training data corpus.

4. Conclusion

The detection and analysis of phoneme boundaries play an essential role in the recognition of speech. But most of the ASR systems have an issue with identifying phoneme boundaries and hence the positions of phonemes. Two hybrid models are developed using Kaldi, one conventional word-based and the other syllable-based, and the results are tabulated and analyzed. The conventional word-based model shows better performance compared to the syllable-based model. But the performance of the syllable-based model was found to be increasing proportionally with the increase in corpus size. The syllable level model is implemented using rule-based syllables for the Malayalam language, generated from a phonetic analyzer, mlphon.

Acknowledgments

The authors express sincere gratitude to Kavya Manohar for sharing the codes of *mlphon*.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability The datasets analyzed in this manuscript are publicly available

References

[1] Babu L B, George A, Sreelakshmi K R and Mary L 2018 Continuous speech recognition system for Malayalam language using Kaldi. In Proc. of Int. Conf. on Emerging Trends and Innovations In Engineering And Technological Research, pp. 1–4

[2] Kurian C and Balakrishnan K 2009 Speech recognition of Malayalam numbers. In Proc. of Nature and Biologically Inspired Computing, (NaBIC), Coimbatore India, pp. 1475–1479

[3] Thennattil J J and Mary L 2016 Phonetic engine for continuous speech in Malayalam. *IETE J. Res.* **62** 679–685

[4] Anand A V, Shobana Devi P, Stephen J and Bhadran V K 2012 Malayalam speech recognition system and its application for visually impaired people. In Proc. of Annual IEEE India Conf. (INDICON), pp. 619–624

[5] Manohar K, Jayan A R and Rajan R 2022 Mlphon: A multifunctional grapheme-phoneme conversion tool using finite state transducers. *IEEE Access* **10** 97555–97575

[6] Lekshmi K R and Elizabeth S 2016 Automatic speech recognition using different neural network architectures - A survey. *Int. J. Comput. Sci. Inf. Technol.* **7**(6) 2422–2427

[7] Mohri M, Pereira F and Riley M 2002 Weighted finite-state transducers in speech recognition. *Comput. Speech Lang.* **16**(1) 69–88

[8] Golob Z, Gros J, Zganec M, Vesnicer B and Dobrisek S 2012 FST-based pronunciation lexicon compression for speech engines regular paper. *Int. J. Adv. Robot. Syst.* **9** 1–9

[9] Baby A, Thomas A L, Nishanthi N L and Consortium T T S, et al 2016 Resources for Indian languages. In Proc. of Text, Speech and Dialogue. CBBLR Workshop

[10] Deekshitha G, Sreelakshmi K R, Babu B P and Mary L 2018 Development of spoken story database in Malayalam language. In Proc. 4th Int. Conf. on Electrical Energy Systems (ICEES), pp. 530–533

[11] Lekshmi K R and Sherly E 2021 An ASR system for Malayalam short stories using deep neural network in Kaldi. *Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, pp. 972–979