Real time Translation of Discrete Sinhala Speech to Unicode Text

M.K.H. Gunasekara and R.G.N. Meegama

Department of Computer Science
Faculty of Applied Sciences,University of Sri Jayewardenepura
Gangodawila, Nugegoda, Sri Lanka

mkh.gunasekara@gmail.com, rgn@sci.sjp.ac.lk

Abstract—**This paper presents a methodology to translate discrete Sinhala speech to Sinhala Unicode text in real time. Initially, the Hidden Markov Model and the associated Hidden Markov Toolkit (HTK) is used as the speech recognizer. While real time decoding is obtained bythe Julius decoder a three-states Bakis HMM topology is used to build the acoustic model. The normalized Mel frequency cepstral coefficients with zeroth coefficient as feature vector is used to recognize speech. Although a single person is used during the training session, an average accuracy of 95% is obtained for both speaker dependent and speaker independent speech recognition. Performance evaluation shows the capabilities of the proposed system to convert discrete Sinhala speech to Sinhala Unicode in both quiet and noisy environments.**

Keywords-Automatic Speech recognition, Hidden Markov Models, Sinhala speech

I. Introduction

In humans, speech is considered as the most natural and the most powerful method of communication. Spoken language has the unique property that it is naturally learned as part of human development. Speech recognition is the process of converting an acoustic signal into its corresponding word sequence. Its applications span fields such as command and control, information retrieval, speaker identification, etc. Hence, applications that are capable of recognizing speech is rapidly gaining ground in the arena of language processing research during the past decade.

As of today, the study of speech recognition has matured to a level where an automatic speech recognition (ASR) routine can be implemented successfully for many languages spoken in the world. Also, speech recognition systems (SRS) are becoming increasingly popular due to their high speech based interactions allowing developers to use such technology to improve users' experience in applications.

The size of a vocabulary of SRSs affect the accuracy of the system. Such vocabularies can be categorize as small (less than 15 words), medium (less than 50 words), large (thousands of words), very large (tens of thousands of words) and open vocabulary, which has no limitations.SRSs are generally divided into two categories, namely, discrete and continuous. Discrete SRSs require the user to leave a brief pause between each spoken word whereas continuous SRSs enable the user to speak continuously just as during a natural flow of words where gaps and pauses between spoken words are not maintained .

Many researchers have attempted to develop ASR, including approaches based on Artificial Intelligence, Support Vector Machines (SVM), Hidden Markov Models (HMM), Artificial Neural Networks (ANN) and Deep Learning algorithms [15]. Out of these numerous techniques,HMM is the most popular statistical approach to automate speech recognition [36].

The HMM, introduced in the 1960's, creates stochastic models from an unknown utterance and compares the probability that the unknown utterance is generated by each model in order to recognize an unknown spoken word. HMMs are used in speech recognitionnot only because a speech signal can be viewed as a piecewise stationary signal but also HMMs can be trained automatically and at the same time computationally feasible.

Sinhala is one of the native national languages in Sri Lanka, a nation island in Asia located near the southern tip of India. The language belongs to the Indo-Aryan branch of the Indo-European languages and contains 40 segmental phonemes, including 14 vowels and 26 consonants [2].

In human speech, a phoneme is the smallest structural element that distinguishes significant units in language such as words or morphemes. Phonemes are cognitive abstractions or categorizations of them, but are not physical segments themselves. Different languages have different number of phonemes, for example, the English language has 44 phonemes including 20 vowels phonemes and 24 consonant phonemes approximately [1]. The Sinhala language contains 40 segmental phonemes, including 14 vowels and 26 consonants [2].

The main objective of this research is to design and develop a mechanism to translate Sinhala speech to Sinhala Unicode text in real time. The major achievements of the proposed project are:

- Ability to recognize Sinhala words in speaker independent scope
- Ability to recognize spoken Sinhala words in a noisy environment.

A. Recent work

HMMs have been used to recognize continuous Sinhala speech in [28] in a speaker dependent environment where an accuracy of 75.74% in recognizing sentences and 96.14% in recognizing words is reported. The training data sets are recorded with 16 KHz frequency and the HTK is used to model the HMMs.

A speaker independent Sinhala speech recognizer for voice dialing is presented in [29] using HMMs to give commands forVoIP applications. Data for the training sessions are obtained by recording speech at a 16 KHz frequency while the HTK is used to model the HMMs. A filtering algorithm embedded into the methodology shows a 82.19% and 87.37% recognition rates in noisy and quiet environments, respectively.

II. METHODOLOGY

The aim of the speech recognition engine (SRE) is to find a sequence of words $W = w_1, w_2, \dots . w_n$ with the maximum likelihood probability $P(W/A)$ for a given acoustic observation$A = A_1, A_2, \dots . A_n$.

There are extremely large number of possible word sequences in a natural language and an enormous range of variation in the acoustic signals produced when different speakers pronounce the same sequence of words. As $P(W/A)$cannot be calculated directly for such a problem, the Bayes' rule is used to break the problem up into two components as follows:

$$P(W/A) = P(W)\frac{P(A/W)}{P(A)}$$

The SRE must model the two probability distributions to calculate $P(A/W)$and $P(W)$. Furthermore, acoustic processing is used to extract the features from the speech waveform which required during recognition. The SRE usually require two essential components in order to recognize speech: acoustic and language model. The acoustic model matches word sequences with similar acoustic properties to an observed speech input. Such acoustic models are created by compiling audio recordings of speech and their transcriptions into statistical representations of sounds. Conversely, a language model gives the probabilities of sequences of words.

In this research, an HMM is used to create the acoustic models using HTK [33, 35] while the language model is created using Grammar.

A. Data Collection

A speech corpus is a vital necessity when developing a speech recognition project. In order to build such a corpus, the recordings are carried out by a native speaker of Sinhala language using the Audacity software [32] with a sampling frequency of 22050 Hz, mono channel. Due to the complexity of spoken Sinhala language, only a medium size vocabulary, containing 50 words, is considered in this research. Subsequently, the recorded files are saved in 16 bit PCM WAV format. Using 100 sentences as training data, recordings are done in a quite environment in which words are assigned to a sentence using a random number generator.

a) Sinhala Pronunciation Dictionary

A pronunciation dictionary, which lists all the words used for the recordings with their phonetic representations of the SRE, is required to train the HMM to work on a phoneme-based speech recognizer as shown in Table 1.As a proper pronunciation dictionary for the Sinhala language is not available, a Sinhala pronunciation dictionary is created for the training dataset.

TABLE 1.
SINHALA PHONEME REPRESENTATION

Words	Phoneme representation
subha (සුභ)	s uh b er
aayuboewan (ආයුබෝවන්)	aa y uh b ao w ah n
udhaesanak (උදැසනක්)	uw dh ae s er n ah k
raathriyak (රාත්‍රියක්)	r aa th r iy y ah k
dhahawalak (දහවලක්)	dh ah hh ah w er l ah k
sAndhAyaawak (සැන්දැයාවක්)	s ah n dh y aa w ah k
ammaa (අම්මා)	ah m m aa
mahathmayaa (මහත්මයා)	m ah hh ah th m er y aa
mahathmiya (මහත්මිය)	m aa hh ah th m iy y er
menawiya (මෙනවිය)	m eh n er w iy y er
yahaLuwaa (යහළුවා)	y ah h ah l uw w aa
thaaththaa (තාත්තා)	th aa th th aa
akkaa (අක්කා)	aa k k aa
ayiyaa (අයියා)	aa ih y aa
oyaa (ඔයා)	ow y aa
gedhara (ගෙදර)	g eh dh er r er
wAdata (වැඩට)	w ae d er t er
saadhayata(සාදයට)	s aa dh er y er t er
welaawata (වෙලාවට)	w ey l aa w er t er
paasAlata (පාසැලට)	p aa s ae l er t er
mama (මම)	m ah m er
ennam (එන්නම්)	eh n n ah m
yannam (යන්නම්)	y ah n n ah m
enawaa (එනවා)	eh n er w aa
yanawaa (යනවා)	y ah n er w aa
kawudha (කවුද)	k ah uw dh er
kohomadha (කොහොමද)	k ow h ow m er dh er
thorathuru (තොරතුරු)	th ow r er th uw r uw
dhawasa (දවස)	dh ah w er s er
jiwithaya (ජීවිතය)	jh iy w iy th er y er

B. Feature Extraction

The speech recognizer needs to differentiate between different kinds of phonemes, the way a human being perceives it. The same word can be spoken by the same person differently at different occasions. As such, the recognizer should be able to identify the exact word though spoken differently. On the other hand, two words which sounds different are perceived as different. It must be noted that sounds have common features even when identical sounds are produced by different speakers. Therefore, a good feature extractor should recognize these features for further analysis and processing during speech recognition.

The feature analysis component of an ASR system plays a crucial role in the overall performance of the system. Several techniques have been developed so far for solving this problem such as the linear predictive cepstral coefficients (LPCC) [3-7] perceptual linear predictive coefficients (PLP) [8], mel-frequency cepstral coefficients (MFCC) [9-11], relative spectra filtering of log domain coefficients (RASTA) [12] and integrated phoneme subspace (IPS) [13].

In literature, several algorithm, such as the principal component analysis (PCA), linear discriminant analysis (LDA) and independent component analysis (ICA), are reported to have been applied on features to increase the performance of the SRE. In this project, LDA is used to transform speech data in the time-frequency domain as it shows better performance than combined linear discriminants in both temporal and spectral domains [14].

MFCCs are widely applied in ASRs and speaker recognition tasks [30]. Therefore, MFCCs are used in this project as features because the Julius decoder is able to support MFCCs [27]. The block diagram of the MFCC processor is shown in Fig. 1.

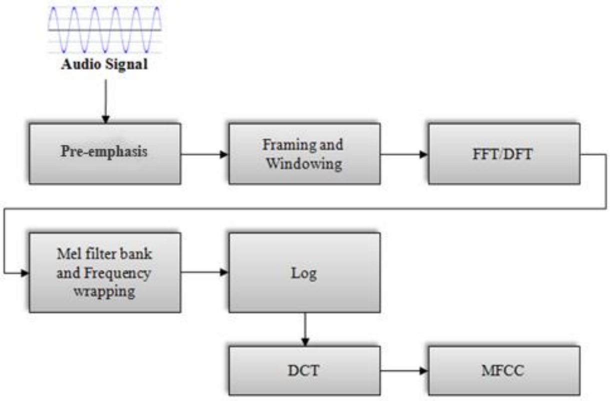


Fig. 1. Components of the MFCC processor.

MFCCs are much more accurate than time domain features [9]. First order regression coefficients (delta coefficients) and second order regression coefficients (acceleration coefficients) are proposed in [31] to add dynamic information to the static cepstral features. The use of about 20 MFCC coefficients is common in ASR although 10-12 coefficients are often considered to be sufficient for coding speech [10]. Normalized MFCC feature vectors and zeroth coefficient are robust to recognition in noise conditions [11]. Therefore, normalized MFCCs, delta coefficients, acceleration coefficients and zeroth coefficient are used in this project as features having 12 cepstral parameters giving a feature count of 39 as depicted in Table II.

TABLE II.
MFCC FEATURE VECTORS

Feature type	Count
Normalized Cepstral Coefficients	12
Delta Cepstral Coefficients	12
Acceleration Cepstral Coefficients	12
Zero Coefficient	1
Delta Zero Coefficient	1
Acceleration Zero Coefficient	1
Total	39

C. Acoustic Modeling

Acoustic modeling of speech is the process of establishing statistical representations for the feature vector sequences computed from the speech waveform. This step models the relationship between the audio signal and the phonetic units in the language. Plenty of technologies have been evolved to do such acoustic modeling such as HMM, ANN [16],[17], ANN-HMM hybrid [18],[19],[20],[21], SVM-HMM hybrid [22] and deep learning[23],[24],[25].

A major challenge in using HMMs is that the topology has to be determined prior to the training and remains fixed during the training phase. The practice is to use left- to–right topology with three or five states to represent a unit (phoneme). A three-state Bakis HMM topology is used in this research as given in Fig. 2.

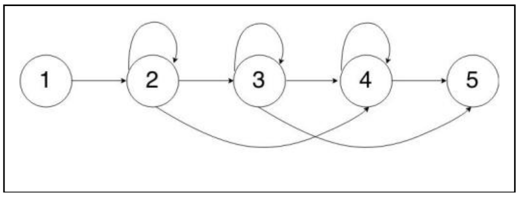


Fig.1. HMM Bakis Topology

D. Language Modeling

In speech recognition, the task of any language model (LM) is to determine a probability for a word given the word history. There are two types of languages models, namely, grammar-based LM and statistical-based LM. SREs with grammar-based LM give less word error rate than SREs having statistical-based LM [26].

The present research utilized a grammar-based LM containing 30 words, categorized into 11 groups, and their probability of occurrences in a given sequence. A recognition grammar essentially defines constraints on which an SRE can expect as input. Because grammar rules for spoken Sinhala are not available, it is modeled as words.

E. Decoding

The Julius decoder [34] is used for real time decoding of spoken words to which the acoustic model, language model and the pronunciation dictionary are taken as inputs as given in Fig. 3.

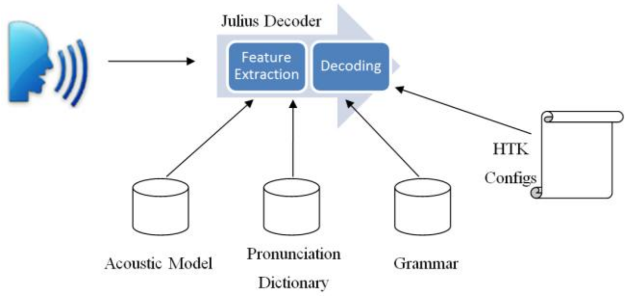


Fig. 3. Architecture of the speech recognition engine.

F. Unicode converter

The range of Sinhala Unicode text is between 0D80 and 0DFF. The conversion into Sinhala Unicode is done by using 'Iskoolapota' font using a java script.

III. Design and Implementation

The information flow of the proposed system is shown in Fig. 4.

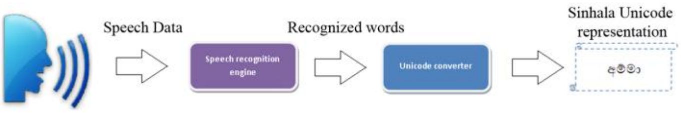
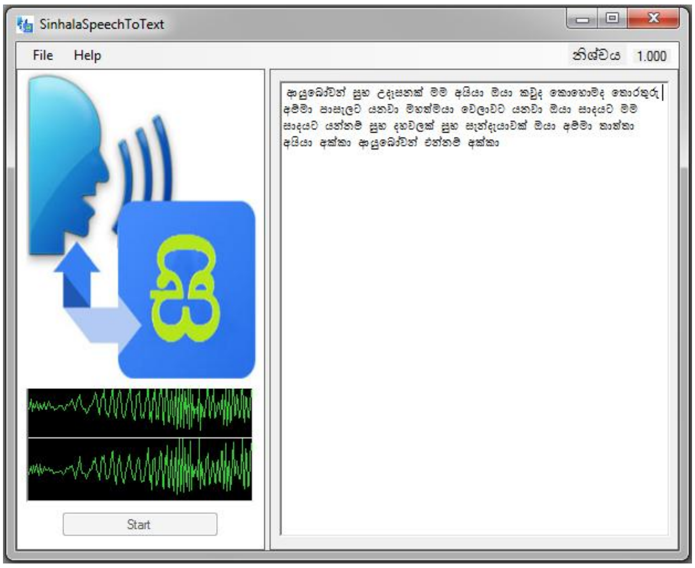


Fig. 4. Pipeline of tasks of the translator.

A. Desktop Application

A desktop application is implemented by integrating the SRE and the Unicode converter using C# to recognize Sinhala speech in real time. The resulting graphical user interface of the application is shown in Fig. 5 from which users are able to copy and paste the translated Sinhala Unicode text to any document processing application as well.



ආයුබෝවන් සුභ උදෑසනක් මම අයියා ඔයා කවුද කොහොමද තොරතුරු
අම්මා පාසැලට යනවා මහත්මයා වෙලාවට යනවා ඔයා පාදයට මම
පාදයට යන්නම් සුභ දහවලක් සුභ සැන්දෑයාවක් ඔයා අම්මා තාත්තා
අයියා අක්කා ආයුබෝවන් එන්නම් අක්කා

Fig.2. User interface of the application.

IV. Results

The proposed system managed to recognize discrete Sinhala speech of different speakers' as well. Therefore, performance evaluation of the system is carried out to verify speaker dependency of the translation using a medium sized vocabulary.

Three speakers (1 male and 2 females) volunteered to test the application for which the results are given in Table III.

The accuracy of translations is computed as follows:

$$Accuracy = \frac{number\ of\ recognized\ words}{total\ number\ of\ words\ spoken} \times 100\,\%$$

TABLE III.
PEFORMANCE EVALUATION FOR SPEAKER DEPENDENCY

Speaker	Accuracy in a quiet environment	Accuracy in a noisy environment
Speaker 1 (male)	94.96%	88.62%
Speaker 2 (female)	97.17%	90.01%
Speaker 3 (female)	97.33%	91.32%
Speaker 4 (trainer)	98.08%	94.80%

A 30 dB of noise is present at the quiet environment whereas a 70 dB of noise is detected at the noisy environment.

V. Conclusion

The proposed research is carried out for a medium sized vocabulary. Although the SRE is trained for a single speaker, recognition rate of speech of different speakers is satisfactory. The reasons for such satisfactory results can be attributed to the following:

a) Usage of 22050 Hz sampling frequency to training speech data

b) The acceleration coefficients, delta coefficients, zero mean static coefficients (normalized) and 0th cepstral coefficients are used as MFCC features.

c) The Bakis topology is used in the HMMto develop the acoustic model.

New technologies such as deep learning and artificial neural networks can be used to create the acoustic model. Extending the research to recognize and translate continuous Sinhala speech to Unicode will be a challenging task. Such an application, if implemented for a mobile phone, will be very much useful for the academic as well as publishing community.

.

VI. REFERENCES

[1] Adriana Vizental, Phonetics and Phonology.: "Aurel Vlaicu" University of Arad, 2010.

[2] W.S. Karunatillake, An Introduction to Spoken Sinhala. Colombo: M.D. Gunasena & Co. ltd., 2004.

[3] S.L. Hanauer and B.S. Atal, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," The journal of the acoustical society of america, pp. 637-655, 1971.

[4] F. Itakura and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," Electronics and Communications in Japan, pp. 36-43, 1970.

[5] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," Acoustics, Speech and Signal Processing, IEEE Transactions on, pp. 57-72, 1975.

[6] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," Acoustics, Speech and Signal Processing, IEEE Transactions on, pp. 336-349, 1979.

[7] Cini Kurian, "A Review on Technological Development of Automatic Speech Recognition," International Journal of Soft Computing and Engineering, pp. 80-86, 2014.

[8] H Hermansky, "Perceptual linear predictive(PLP) analysis of speech," The journal of the acoustical society of america, pp. 1738-1752, 1990.

[9] Lei Xie and Zhi-Qiang Liu, "A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation," in International Conference on Machine Learning and Computing,Dalian, 2006, pp. 4359 - 4364.

[10] A Hagen, D.A Connors, and B.L Pellom, "The Analysis and Design of Architecture Systems for Speech Recognition on Modern Handheld-Computing Devices," in First IEEE/ACM/IFIP international conference on hardware/software design and system synthesis, California, 2003, pp. 65-70.

[11] Olli Viikki and Kari Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Communication, pp. 133-147, 1998.

[12] Hynek Hermansky and Nelson Morgan, "RASTA Processing of Speech," IEEE Transactions on Speech and Audio Processing, pp. 578-589, 1994.

[13] Hyunsin Park, Takiguchi Tetsuya, and Ariki Yasuo, "Integrated Phoneme Subspace Method for Speech Feature Extraction," EURASIP Journal on Audio, Speech, and Music Processing 2009, 2009.

[14] O-W Kwon and T-W Lee, "Phoneme recognition using ICA-based feature extraction and transformation," Signal Processing, pp. 1005-1019, 2004.

[15] Li Deng et al., "Recent Advances in Deep Learning for Speech Research at Microsoft," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013, pp. 8604-8608.

[16] Richard P. Lippmann, "Review of Neural Networks for Speech Recognition," in Neural Computation, 1989, pp. 1-38.

[17] J Tebelskis, "Speech Recognition using Neural Networks," Pittsburgh, Phd Thesis 1995.

[18] H Bourland and N Morgan, Connectionist speech recognition: a hybrid approach. Boston: Kluwer Academic, 1994.

[19] Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems.," in Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, 2000, pp. 1635-1638.

[20] T Robinson, M Hochberg, and S Renals, The Use of Recurrent Neural Networks in Continuous Speech Recognition. Norwell: Kluwer Academic Publishers, 1995.

[21] W Reichl and G Ruske, "A hybrid rbf-hmm system for continuous speech recognition," in International Conference on Acoustics,Speech and Signal Processing, Detroit, 1995, pp. 3335–3338.

[22] Aravind Ganapathiraju, J. E Hamaker, and Joseph Picone, "Applications of support vector machines to speech recognition," in Signal Processing, IEEE Transactions on, 2004, pp. 2348-2355.

[23] Geoffrey Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, pp. 82–97, 2012.

[24] Geoffrey Hinton, George E. Dahl, and A. Mohamed, "Acoustic modeling using deep belief networks," Audio, Speech, and Language Processing, IEEE Transactions on, pp. 14-22, 2011.

[25] Li Deng, Dong Yu, and Ossama Abdel-Hamid, "Exploring convolutional neural network structures and optimization techniques for speech recognition.," In INTERSPEECH, pp. 3366-3370, 2013.

[26] Beth Ann Hockey and Manny Rayner, "Comparison of Grammar-Based and Statistical Language Models Trained on the Same Data," in In Proceedings of the AAAI Workshop on Spoken Language Understanding, 2005.

[27] Akinobu Lee, Kiyohiro Shikano, and Tatsuya Kawahara, "Recent progress of open-source LVCSR engine Julius and Japanese model repository," 2004.

[28] Thilini Nadungodage and Ruvan Weerasinghe, "Continuous Sinhala speech recognizer," in In conference on Human Language Technology for Development, Alexandria, Egypt, 2011.

[29] W. G. T. N. Amarasingha and D. D. A. Gamini, "Speaker Independent Sinhala Speech Recognition for Voice Dialling," in ICTer, 2012.

[30] Paul Mermelstein and Davis Steven, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," Acoustics, Speech and Signal Processing, IEEE Transactions on, pp. 357-366, 1980.

[31] S Furui, "Speaker-independent isolated word recognition based on emphasized spectral dynamics," in Proc. ICASSP, 1986.

[32] Audacity. [Online]. http://audacity.sourceforge.net/ (Accessed: 7 January 2015).

[33] University of Cambridge. Hidden Markov Model Toolkit. [Online]. http://htk.eng.cam.ac.uk/ (Accessed: 7 January 2015).

[34] Kyoto University. Julius. [Online].

http://julius.sourceforge.jp/en_index.php (Accessed: 7 January 2015).

[35] Steve Young et al., The HTK book (for HTK version 3.4).: Microsoft Corporation, 2006.

[36] Rabiner, Lawrence. "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE 77, no. 2, pp.257-286,1989.