ISSN No:-2456-2165

Development of Speech Recognition System Based on CMUSphinx for Khmer Language

Channareth Srun[1], Virbora Ny[1], Chea Cheat[2], Sokheang Ching[3], Piseth Ny[4]
Faculty of Electronics
National Polytechnic Institute of Cambodia
Phnom Penh, Cambodia

Abstract:- This paper described the process of creating and testing the offline Khmer speech recognition system. This system was created using CMUSphinx with the noise reduction of training audio database, including 85 speakers and 157 words selected from the Khmer language. To evaluate the speech recognition accuracy, there were 100 Khmer transcripts randomly created from the training dictionary for calculating the word and sentence error rate. The recognition accuracy of Khmer speech recognition can achieve up to 89.91% of word recognition accuracy and 90.02% of sentence recognition accuracy.

Keywords:- *Khmer, phoneme, ASR, CMUSphinx Toolkit, Acoustic Model, Phonetic Dictionary, Language Model.*

I. INTRODUCTION

With the advancement in modern-day technology, the requirement of data downloading and sharing between devices is available in terms of wired and wireless technologies. Along with the evolution of Industrial 4.0, we have noticed a significant change in the electronics device's ability to learn and understand the working process based on artificial intelligence. This lead to a part in which the communication between humans and machines can accomplish using voice command through the help of speech recognition technology. Automatic speech recognition is a branch of audio processing and machine learning technology that could transform the input signal from input devices such as microphones to human-readable formats [13].

In 1971, Allen Newell as the leading researcher conducted the case research on speech recognition. The speech recognition systems come with the problems such as Acoustic, Parametric, Phonemic, Lexical, Sentences, and Semantic. Later on, a research team from Carnegie Mellon University in 1976 leading by Mr.Reddy continuing the research and got the result as speech recognition system including Hearsay, Dragon, Harpy, and Sphinx I/II. For the last 40 years, speech recognition has been using the same method in modeling to find Global Optimization. In the report of making speech recognition model, Hidden Markov Model was used to analyze the model and use Nested Block for computation. Nowadays, speech recognition model tools have been available as open-source software such as HTK, Sphinx, Kaldi, CMU LM Toolkit, and SRILM [1].

CMUSphinx toolkit had the advantage of the efficient algorithms for speech recognition with low-resource devices with the flexible design, which is suitable for creating flexible and high accuracy for Khmer automatic speech recognition system. The main objective of this research is to develop an offline Khmer speech recognition system that could convert Khmer speech into readable text. The trained Khmer speech recognition system could run on the personal computer and embedded system for low power computing.

II. BASIC THEORY

Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the process of converting a speech signal to a sequence of words through an algorithm implemented as a computer program [2]. The ASR is a technology that enables computing devices to convert human spoken words into computer-readable text via microphone or telephone input [6]. As shown in Figure. 1 the task of capturing an acoustic signal by using a microphone or another recording device and then transforming the audio signal to readable text. Building a new and robust ASR system for a new language is a task that required a lot of time, resources and effort [4]. Carnegie Mellon University (CMU) Sphinx speech recognition toolkit is freely available as open-source and currently is one of the most robust speech recognizers in English [5].

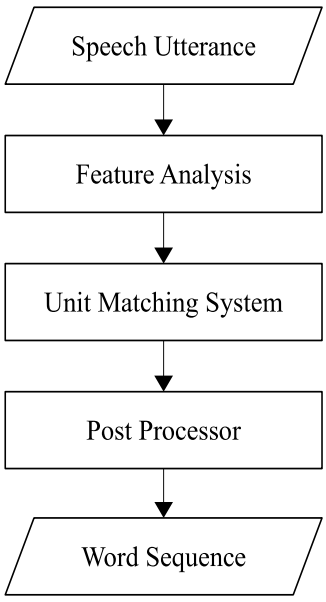

Fig. 1: Structure of automatic speech recognition [6]

ISSN No:-2456-2165

Structure of Speech

Speech is a continuous stream of sound with stable states mixed with dynamically changed. The *phone* is a sequence of states, defined as the classes of sounds and the core of building a word, the acoustic properties of a waveform in each word corresponding to a phone. These conditions had a significant difference depending on many factors such as phone context, speaker, style of speech, and so forth [7].

Speech Recognition Models

Acoustic Model: contains acoustic properties of each Senone, the statistical representation of sounds make up a word, which called *phoneme.*

Phonetic Dictionary: contains the mapping connection of words to phones. Due to only two or three different variants are available for each word, this method is not the most effective but having good practical results. Some other methods are also available for mapping words to phones with complex functional learned with a machine learning algorithm can also be used.

Language Model: restricting the word searching process, the new word got determined from the previously recognized words. An *n-gram* is a commonly used language model, which contains statistics of word sequences and finite-state [7].

Hidden Markov Model

Hidden Markov Models (HMMs) is a statistical Markov model that provides a simple and effective framework for modeling time-varying spectral vector sequences such as speech waveform. Consequently, most present-day large vocabulary continuous speech recognition (LVCSR) systems based on HMMs. Figure. 2 shows the architecture of HMMs speech waveform from the audio source got converted into a sequence of fixed-size acoustic vectors called feature extraction. The decoder later generate the word sequences, which generated the feature vector.

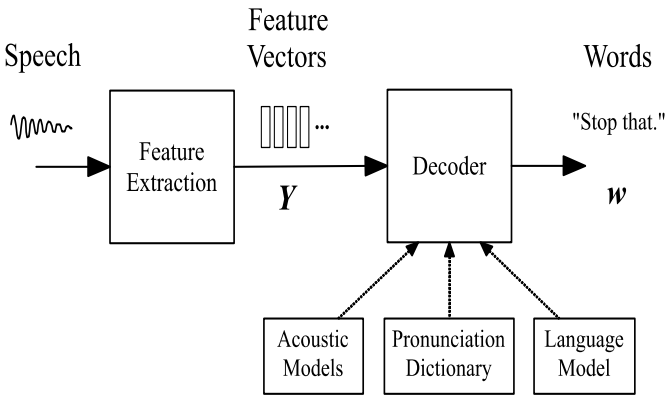

Fig. 2: The HMM-based recognizer's architecture [8]

Feature Extraction: The feature extraction stage provides a small representation of the speech waveform by dividing the features from the dataset into smaller informative feature categories. This form minimizes the loss of information that specifies between words and provide a good match with the distributional assumptions made by the acoustic models [8].

Gaussian Mixture Model

The commonly used extension to the standard HMMs is to model the state-output distribution as a mixture model. Figure. 2 described a single multivariate Gaussian distribution used for modeling the state-output distribution of HMMs. This model assumed that the observed feature vectors are symmetric and unimodal. In practice, speaker, accent, and gender differences would create multiple modes in the data. To solve this problem, the Gaussian mixture model would replace the Gaussian state-output distribution that gives the flexible distributional to the model in asymmetric and multi-modal distributed data [8].

CMUSphinx Toolkit

Figure. 3 described the two main parts of the CMUSphinx automatic speech recognition system. These system blocks included the Model Generation and Recognition block. Model generation block is the part where the acoustic model is generated using the Sphinx training tool from the training database for later use in speech recognition. The recognition block required the acoustic model and the language model for the recognition process [9].

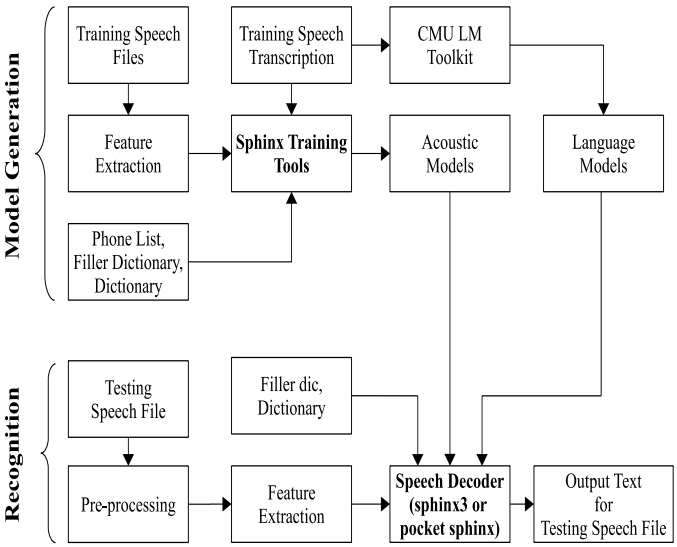

Fig. 3: Block diagram of ASR including model generation and recognition [9]

CMUSphinx is a group of open-source software developed by Carnegie Mellon University and a collection of 20 years of CMU research, including speech recognizer and acoustic model trainer such as:

SphinxTrain: is a set of software used for training script and acoustic models for creating speech recognition model for any language with sufficient acoustic data that can run with CMUSphinx speech recognizer.

Sphinx: is a speech recognizer that can run continuously without depending on the specific speaker that converts speech into readable text using a trained acoustic model created by SphinxTrain based on Hidden Markov Acoustic Model and Gaussian Mixture Model.

PocketSphinx: is a smaller version of Sphinx designed for running on low resource platforms, especially the embedded system that is based on the ARM processor including features such as Fixed-point Arithmetic with

the algorithm which able to compute Gaussian Mixture Model.

III. EXPERIMENTAL METHODS

The Training Database Preparation

In the process of creating the Khmer acoustic model, 157 words in the Khmer language were chosen as corpus that the phoneme based on the English phoneme pronunciation [10]. In this step, 49 phonemes are used as the base of the Khmer word's pronunciation. To collect training data, there are 85 individual speakers recorded with noise-canceling microphones. The duration of each spoken word is 1.5s in a total of 13345 audio files and 5.56 hours of total recording time.

Table I summarize the specifically required audio file format by the CMUSphinx acoustic model toolkit. The SphinxTrain will start the training process throughout the parameters of the sound unit from the sample of speech signal for the model called the training database, which contains information required to extract statistics from the speech in the form of the acoustic model [11].

TABLE I. REQUIRED AUDIO FORMAT FOR TRAINING DATABASE RECORDING

Audio Parameter	Value
Format	WAV
Sampling Rate	16kHz
Bitrate	16bit
Encoding	Little Endian
Channel	Mono

Training The Khmer Acoustic Model

In the training process, a training database is required. In this process of training the Khmer acoustic model, the database is the representation of the speech from the mobile recording channel with a variety of recording conditions. The database has to be split into two parts, the training part, and the testing part. The testing part should be 1/10th of the total data size and suggested not to be more than 4 hours of the recording test data [11]. The database prompts with post-processing will include the files with the structure of the database as follow:

Phonetic Dictionary: contained the phonetic transcription of each transcript in the dictionary file, which let the decoder, knows how to pronounce each word.

Table II described a briefing process covering on Khmer transcripts are converted into Unicode transcripts which contained the resemble phonetic to the Khmer transcript

TABLE II. PHONETIC TRANSCRIPTION

Khmer Transcript	Unicode Transcript	Phonetic Transcription
ប្រទេស	BRATES	B R O T E SS
កម្ពុជា	KAMPOUCHEA	K AE M P AW CH IY AH
ជំរាបសួរ	CHOMREABSUOR	CH AA M R IY B S UW W AH R

Phoneset: contained the list of Phoneset that matched the phones used in the training dictionary file and including the SIL for silence utterance.

Table III listed all the phone sets used in training Khmer acoustic model.

TABLE III. PHONEME USED IN TRAINING THE ACOUSTIC MODEL

A	AA	AE	AH	AI	AO	AW
AY	B	CH	D	E	EA	EI
EH	ER	EY	F	G	H	HH
I	IE	IH	IY	J	K	L
M	N	NG	O	OH	OR	OU
OW	OY	P	R	S	SS	T
TH	U	UW	V	W	Y	Z

Fillers Dictionary: contained the phones that are not linguistic sound included by the language model such as breath, hmm, uh, um, cough, or laugh:

<s> SIL
</s> SIL
<sil> SIL

Fileids: are files contained text while listed all the names of recorded utterance:

speaker1/SUOSDEI_1
speaker2/SUOSDEI_2
.
.
.
speaker85/SUOSDEI_85

Transcription: are files contained text listed the transcription for each audio file

<s> SUOSDEI </s> (speaker1/SUOSDEI_1)
<s> SUOSDEI </s> (speaker2/SUOSDEI_2)
.
.
.
<s> SUOSDEI </s> (speaker85/SUOSDEI_85)

Recording of speech utterance: is the audio file contained in the training database that matched the speech of the word that would be recognized in the speech recognition process. In case of a mismatch, the transcript and the audio data will cause a significant drop in the recognition accuracy [11].

Training Results

During the training process, the decoder computes the acoustic model with the test part of the training database and the reference transcripts to calculate the Word Error Rate (WER) and Sentence Error Rate (SER) of the model.

Table IV shows the results of WER and SER of Khmer acoustic model determined by the decoder

TABLE IV. WER AND SER OF THE ACOUSTIC MODEL TRAINING RESULT

Error Rate Evaluation	Error	Corpus
SENTENCE ERROR RATE	10.0%	(1329/13345)
WORD ERROR RATE	10.1%	(1346/13345)

The decoder proceeded with the calculation and generated the acoustic model's description for the acoustic model in the .align file. Table V shows the percentage correctness of the acoustic model along with error and total recognition accuracy.

TABLE V. THE DECODER CALCULATAION OF ACCURACY

Percentage correct	Error	Accuracy
90.12%	10.09%	89.19%

IV. EXPERIMENT RESULTS

Testing Methods

To determine the speech recognition accuracy through live speech, the calculation of WER could be calculated using the Equation 1.

$$\text{WER} = \frac{I + D + S}{N} \quad (1)$$

The accuracy value can be calculated in Equation 2 for estimating the performance of the decoder.

$$\text{Accuracy} = \frac{N - D - S}{N} \quad (2)$$

Where:

- Insertions (I): are the incorrectly inserted words in the recognized transcript.
- Deletions (D): are the undetected words in the recognized transcription.
- Substitutions (S): the substituted words between reference and recognized transcript.
- Number of words (N): the numbers of counted words in a transcription.

Figure. 4 describes the process of evaluating the Khmer speech recognition accuracy that can be done by randomly created 100 transcripts using the word list from the training dictionary file. Each transcript was tested three times with different testing conditions to get the I, D, S parameters for later calculation for WER. Using WER from each sentence, the recognition accuracy can be determined.

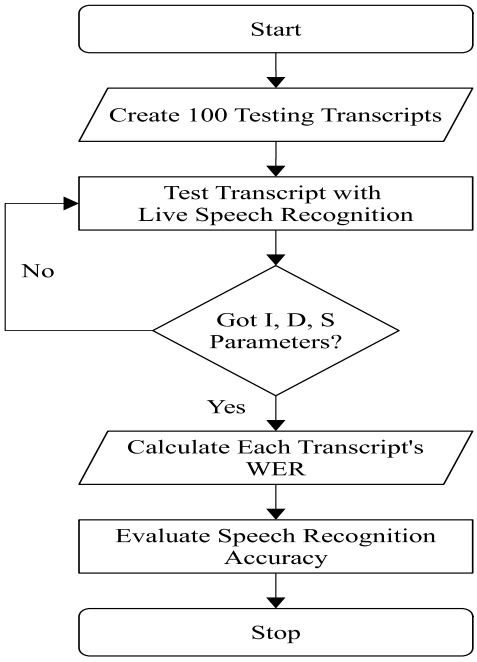

Fig. 4: Process of evaluating Khmer speech recognition accuracy

Figure. 5 shows the live speech recognition is operated by running the PocketSphinx with the Khmer acoustic model on a personal computer. The later process is comparing the recognition result with the written transcripts to get the I, D, S parameters for further estimating the recognition accuracy. "SUOSDEI BRATES KAMPOUCHEA" is the output text converted from a speech by the decoder.

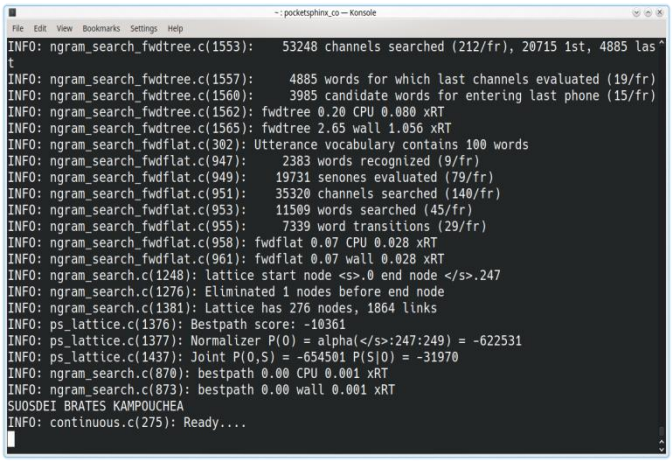

Fig. 5: Live speech transcript testing

Table VI describes a briefing process about analysis with the written transcript to the live speech recognized words. The result of the WER value are calculated and determined the recognition accuracy of each transcript.

TABLE VI. BRIEFING TASK OF WORD ERROR RATE CALCULATION

Khmer Transcript	Unicode Transcript	Speech Recognition	I	D	S	N	WER	Accuracy
សួស្តីប្រទេសកម្ពុជា	SUOSDEI BRATES KAMPOUCHEA	SUOSDEI BRATES KAMPOUCHEA	0	0	0	3	0.00	100.00
ភាសាខ្មែរ	PHEASA KHMER	PHEASA KHMER	0	0	0	2	0.00	100.00
ខ្ញុំមកពីប្រទេសកម្ពុជា	KHNHOM MOK PI BRATES KAMPOUCHEA	KHNHOM MOK PEAK BRATES KAMPOUCHEA	0	0	1	5	20.00	80.00

Figure. 6 shows the difference between each Khmer acoustic model sentence recognition accuracy through each different version. There are four different versions of the Khmer acoustic model created using a noise-canceling training database. Each version went through the training process with different database sizes and optimization parameters.

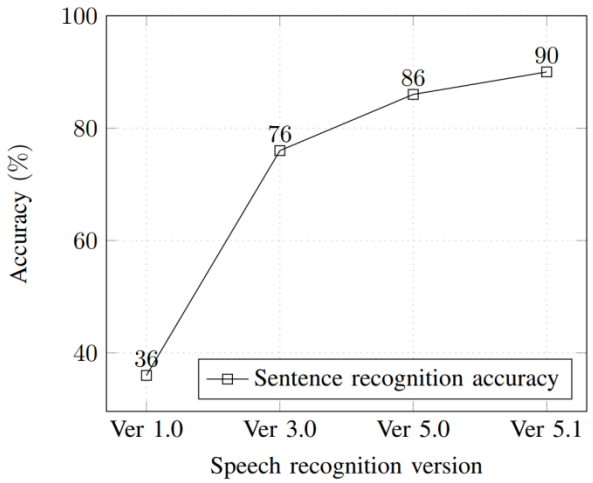

Fig. 6: Khmer speech recognition sentence recognition accuracy

Figure. 7 shows the difference between each Khmer acoustic model word recognition accuracy. Noticeably the word recognition accuracy of the acoustic model is slightly lower than the sentence recognition accuracy shown in Table IV due to higher corpus error in the acoustic model training process.

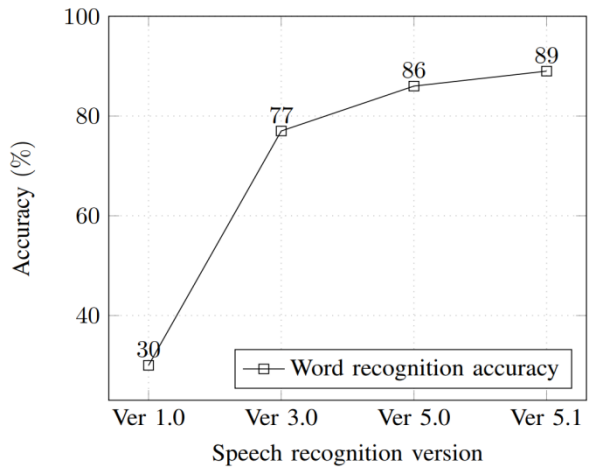

Fig. 7: Khmer speech recognition word recognition accuracy

V. CONCLUSION

In conclusion, using the noise reduction audio database as the recognition accuracy improvement method, the Khmer automatic speech recognition can convert the speech audio into words sequence from live speech and digital audio files with a total recognition accuracy of up to 90.12%. Due to the low system resource requirement to run on a platform through the CMUSphinx toolkit, the Khmer ASR is also suitable for embedded systems with better performance and recognition accuracy [12].

VI. ABBREVIATION AND ACRONYMS

ASR – Automatic Speech Recognition
CMU – Carnegie Mellon University
GMM – Gaussian Mixture Model
HMM – Hidden Markov Model
LVCSR – Large Vocabulary Continuous Speech Recognition
PDF – Probability Density Function
PMF – Probability Mass Function
SER – Sentence Error Rate
SIL – Silence
WER – Word Error Rate

REFERENCES

[1] J. Huang, "A Historical Perspective of Speech Recognition," *Communication of ACM*, vol. 57, no. 1, pp. 94-103, 2014.

[2] S. M.A.Anusuya, "Speech Recognition by Machine: A Review," *International Journal of Computer Science and Information Security*, vol. 6, no. 3, 2009.

[3] R. F. R. H. Hamdan Prakoso, "Indonesian Automatic Speech Recognition System Using CMUSphinx," *International Symposium on Electronics and Smart Devices (ISESD)*, pp. 277-288, 1993.

[4] V. Stouten, "Robust Automatic Speech Recognition in Time-Varying Environments," 2006.

[5] M. N. H.Satori, "Arabic Speech Recognition System Based on CMUSphinx," p. 2007.

[6] R. Prakoso, "Indonesian Automatic Speech Recognition System Using CMUSphinx Toolkit and Limited Dataset," *International Symposium on Electronics and Smart Devices (ISESD)*, pp. 283-286, 2016.

[7] C. M. University, "Basic Concepts of Speech Recognition," [Online]. Available: https://cmusphinx.github.io/wiki/tutorialconcepts/. [Accessed 4 December 2020].

[8] S. Gales, "The Application of Hidden Markov Models in Speech Recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195-304, 2008.

[9] P. A. V. M. J. N. V. S. I. B. B. H. S. J. E. S. M., "Transcription of Telugu TV News using ASR," *International Conference on Advances in Computing, Communications and Informatics (ICCACCI)*, 2015.

[10] P. V. Ferdiansyah, "Indonesian automatic speech recognition system using English-based acoustic model," *Proc. of International Conference Electrical Engineering and Informatic (ICEEI)*, pp. 1-4, 2011.

[11] C. M. University, "Training an acoustic model for CMUSphinx," [Online]. Available: https://cmusphinx.github.io/wiki/tutorialam/. [Accessed 6 August 2021].

[12] Microsoft, "Evaluate and Improve Custom SPeech Accuracy," [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data. [Accessed 5 August 2021].

[13] B. C. Kurian, "Speech recognition of Malayalam numbers," *International Journal of Computer Science and Information Security*, vol. 6, no. 3, 2009.