Enhancing Speech Recognition In Developing Language Learning Systems For Low Cost Androids

Akshay Jayakumar[1], Meera Raghunath[2], Sakthipriya M.S[2],
Akhila S[2], Anuja Sadanandan[2], Prema Nedungadi[1]
1. Amrita CREATE
2. Department of Computer Science and Engineering
Amrita VishwaVidyapeetham
Amritapuri, Kerala, India
akshayjayakumar@am.amrita.edu,mssakthipriya@gmail.com, meerakrishna93@yahoo.co.in, prema@amrita.edu

***Abstract*—Learning to read correctly is a key requirement of language learning. In rural India, due to lack of teachers and technology, tablets offer a creative and motivating learning environment. Tablet technology has the advantage of mobility, allowing users to learn at their own pace and convenience. However, the non-availability of electricity and Internet can be unique challenges. At Amrita CREATE, language-learning solutions have been developed for students to learn and read on the tablets. It uses advanced speech recognition technique to provide feedback and intervention. Proposed system is unique in its ability to evaluate words and phrases and corrects the learner as they articulate the sentence. This system works without Internet and on the lower processing power of android tablets. Silence detection and multiple-synchronized recognition have been introduced in this paper which greatly enhance the ability to provide feedback to the user in real-time. The combination of the two helps in achieving successful recognition of longer and continuous sentence.**

Keywords—SpeechRecognition; Real-Time Error Detection; Real-Time Feedback; Technology Enhanced Learning; Low Cost Tablets; Rural Education; Pocket Sphinx; Silence Detection; Multiple Synchronized Recognition.

I. INTRODUCTION

Aiming to provide literacy and awareness among the rural population in India, Amrita CREATE has developed a unique literacy program called Amrita RITE. This is essentially a combination of Low Cost Tablet (LCT) technology and pedagogy [1], which has been successful in delivering quality contents there by giving access to free education across rural India. Amrita RITE has revolutionized the way technology collaborates with education. It is a major contributor towards eradication of literacy in India [1] by enhancing reading and writing in multi-grade classrooms in rural India [1]. India has as many as 22 major languages. Catering to such a diverse community is indeed a challenge. Amrita RITE provides a solution that can be applied to any Indian languages. Learning to read and write in any one language is the underlying idea of literacy in India. However practical situation demand and require, a fair knowledge in international language, English and official language, Hindi, for effective communication. In order to achieve this, a new approach has been developed that allows learners to learn through reading and writing.

Technology enhanced learning has dramatically evolved in the recent past. Newer learning platforms like e-learning and m-learning that adopt various technologies into the classroom settings are introduced. It helps to improve teacher-learner interaction, knowledge base expansion and provide easier methods to analyze progress and performances [2]. Computer Assisted Language Learning (CALL) talks about the potential to replace the traditional learning methodology, since it works in more flexible environments and offers better options in relation to the learner/teacher needs, interests and abilities [3], [4]. By redefining this concept, Amrita CREATE has been successful in establishing Tablet Enhanced Learning in Rural villages across India [1]. The system introduced here is yet another addition to it. This system is capable of handling language learning by providing a platform to improve pronunciation.

The system introduced here for language learning uses speech recognition and is developed in the Android Open Source Project platform. Pocket sphinx speech engine is implemented for speech recognition. Key focus of research is to provide a word-by-word feedback in real-time. Most of the speech recognition software is able to provide recognition result only after articulating the entire sentence. But this paper describes how silence-detection and multiple-synchronized recognition can be integrated together to achieve word-by-word recognition and feedback in real-time.

This system listens and evaluates speech as it is spoken. Any mispronounced word(s) in the articulated speech is recognized in real-time and the learner is made to correct it and proceed. Unlike other language learning tools, this system allows the learner to correct and continue articulation rather than to start from the beginning. This is found to be very motivational for young learners and also helps greatly in improving the pronunciation. In this combination the recognition process is independent of the length of the articulated speech, thereby improving its efficiency.

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

II. RELATED WORKS

Speech recognition technique has been implemented in multiple educational tools with a focus on enhancing spoken language skills. Such tools are developed to aid native/ non-native speakers to perfect their skills by improving vocabulary and pronunciation. In a particular implementation, common mistake usually made by the non-native speakers are identified and a feedback with visual reference to correct the pronunciation is provided [5]. Yet another implementation recognizes difficult words when the user stumbles while reading out aloud to the system. More complex parameters related to speech such as pitch, duration and score of spoken phrases is used to identify mispronounced words in a speech [6]. It compares proper speech with user spoken speech, identifies mispronunciation with specific deviation parameters. Phone level comparison is also done to assess the pronunciation mistakes by comparing non-native speech with a native speech [10].

The focus of the research described in this paper is to develop a personalized-learning system that allows real-time feedback in correcting pronunciation mistakes and advancing to the further stages of speech learning. Another major focus is to make the recognition process independent of the length of the articulated speech. This greatly enhances the recognition efficiency while working with long and continuous speech.

III. PROPOSED METHOD

Providing real-time evaluation and feedback is very important in any learning tool especially in the one that is used for improving speech and pronunciation. Research suggests most educational tools fail to achieve this. In many systems that implement speech recognition, the recognition results are obtained at the end. The articulated speech is completely recorded/ captured in the primary stage. Subsequently, the captured speech is subjected to the process of recognition. After this stage, the recognition results are obtained. This recognition results are then evaluated to provide a feedback to the learner. Often, a small error in the word pronunciation would affect recognition and this could be identified only the end. In order to correct the mistake, learners have to articulate the entire speech and wait for the results. This is found to be less motivating as the whole speech/sentence has to be repeated again. Yet another drawback with such systems is the length of the articulated speech impacts the speed of recognition. More the length of the articulated speech, the more time it takes to process and provide the results. Sometimes, the length of the articulated speech is significantly large to create insufficient memory exceptions and results in fatal errors especially in low processing / low memory systems. Thus in any reliable educational tool, such shortcomings are considered less desirable.

The system mentioned in this paper is a functional improvement to the ones mentioned above. This system works on the fundamental assumption that every word in a sentence is followed by a pause/ silence. Proposed system allows learners to correct mistakes and advance as and when they occur, rather than start articulating the speech from the beginning. In this approach, the system continuously listens to articulated speech and splits it into different snippets by analysing the silence in it. Every snippet is subjected to the recognition process using a unique instance of the speech recognition and is evaluated separately. This methodology allows the articulation and recognition to happen simultaneously. Since the recognition is independent of the entire articulated speech, any mispronounced word is identified instantaneously, facilitating immediate feedback to the users in real-time.

IV. METHODOLOGY

The system discussed in this paper is implemented as an Android Application running on a tablet. In this application, the speech improvement / learning process is themed as a conversation between two characters. Once user select from a list of conversations they are prompted to speak a word or a sentence displayed on the screen for which evaluation is performed. As the process commences, the system continuously listens to the audio input from the user. User articulated speech is the input to the system. All along the articulation, the silence detection module detects the pause/ silence between the words and each word is captured separately as an audio snippet. Each such snippet is processed by the speech engine decoder which recognizes the word and transcribes it. This is then returned as the result of the recognition. This result is further evaluated against a predefined library of words and the feedback is provided to the user all in real-time. This design is designed to recognize even longer sentences without compromising on the recognition / feed-back accuracy. Even when the articulation is underway, visual feedbacks are provided to the learner. Green highlighted text indicates the proper pronunciation of the words and red indicates the mispronounced words (Fig. 1). This gives the learner with a choice to correct the mistakes and continue without the need of articulating the entire speech from the beginning.

Fig. 1. Real-time recognition results as feedback .

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

The working of this system is divided into four phases: (1) Silence detection (2) Speech recognition (3) Evaluation/ Comparison (4) Feedback presentation. Interaction between all these modules is depicted in Fig 2.

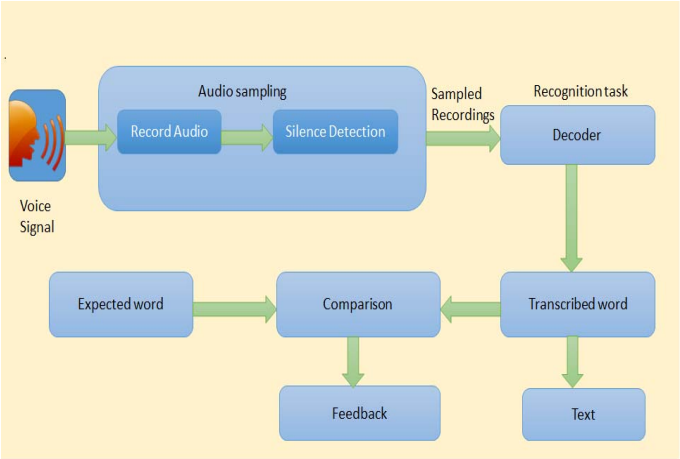

Fig. 2. Overall architectureof the system.

A. *Speech Recognition*

Speech recognition is one of the most important phases. All the subsequent steps depend on the speed and accuracy of this phase. The speech recognition system transforms the speech signals into a sequence of text. For speech recognition, Pocket Sphinx speech recognition engine is implemented in the system. Pocket sphinx is a good choice for real-time applications because of its accuracy. It is a speaker-independent and continuous speech recognition system. Pocket Sphinx requires a dictionary, an acoustic model and a language model to recognize the speech. The language model used to configure the speech decoder has a finite set of vocabulary. The words in this vocabulary are the ones that is been taught.

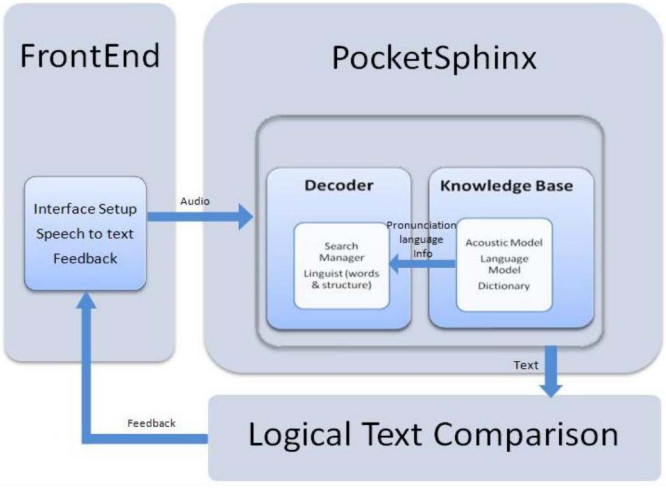

Fig. 3. Component level diagram of PocketSphinx

The device hardware (microphone and sound-card) converts sounds (analogue waves) of the speech into a digital format. Front End is responsible of carrying out digital signal processing on the input digital audio signal. Features or characteristic attributes in accordance to the input sound is produced by Front End. Features are passed on to the next stage where the Decoder uses the Search manager to perform a search using different algorithms (Graph Search). This is based on the criteria predefined such as grammar, knowledge from the dictionary, acoustic model and the language model. The software acoustic model generates phonemes associated with these sounds. The language model compares phonemes to words using its built-in dictionary. Decoder result usually contains multiple best results. It is the speech engine that decides the best spoken word/ close match based on higher confidence value and returns recognition results. Major constraints affecting the accuracy of speech recognition are following:

1. Accuracy of the Acoustic model. Acoustic model should often contain / trained with large amount of acoustic data.

2. Language model and dictionary: More words and sentences we include in the model, the more accurate will be the recognition engine.

3. Noise: Anything causing interference or acts as disturbance during speech articulation is considered as noise. E.g. surrounding sounds, electronic interferences etc. These are basically unwanted in information in the articulated speech

B. *Silence Detection*

Every word in a sentence is followed by a slight pause/ silence. When user articulates a sentence, the system splits it into words by identifying the silence in it. This is done based on the assumption that non–native speakers / beginners articulate words in an optimal rate. The system starts recording when a voice presence is detected and stops when silence is detected. If the amplitude of the sound wave is in a comparable range of zero and a specific threshold value it signifies the presence of a sound. If the sound wave is out of bound of the above specified range, it is considered as a silence. When the silence is detected, the recording stops and a new snippet of audio will be created. This process continues until articulation is complete. Algorithm describing the logic of silence detection and audio snippet creation is given below. Newly created audio snippet is further subjected to the recognition process.

```
WHILE Audio Available:
Mark Recording Flag

IF Amplitude within boundaries
        Voice detected
        Start Recording
        Create Audio Snippet
ELSE
        Silence detected
```

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

```
        Stop Recording
        Save Audio Snippet
        Mark Recording Flag
ENDIF

IF Recording stopped
        Start Recognition Task
ENDIF
```

C. *Multiple Synchronized Recognition*

The system records the speech and splits it into different recordings using Silence Detection. This generates multiple audio snippets corresponding to the words in the sentence. When a new audio snippet is generated, it is processed by one instance of the speech decoder for recognition. Each instance of the speech decoder is considered as a separate thread that executes it. Likewise, all subsequent new snippets are processed in a separate thread running a decoder, which is created during runtime.

A thread pool executer is responsible for the parallel and synchronized execution of these threads. This ensures that correct feedback is given to the user in the order of appearance of the words in the articulated speech. Furthermore, a queue system is implemented to store all active speech decoders. When a new snippet of audio is available, the system checks the queue for an active and free instance of the decoder to allocate for processing. A new decoder instance is created if decoders are not available from the queue. Whenever a decoder instance completes its recognition task, it is stored back in the queue. The living-time parameter associated with each decoder instance determines how long it is maintained in the queue. Decoder instance which exceeds the living-time are destroyed to free up the memory.

In multiple synchronized recognition, each instance of the recognition is completed quickly as it works on one word at a time rather than on a whole sentence. Even when the articulated speech is long, the speech is split into multiple snippets, word-by-word, and multiple recognition instance completes the whole process. This makes the entire recognition process independent of the length of the articulated speech.

```
INITIALIZE Queue
NEW Decoder instance
PUSH Decoder instance to Queue

IF Queue not Empty
        Decoder = POP Queue
        Assign Audio Snippet to Decoder
        RETURN Hypothesis from Decoder
ENDIF

IF Recognition is Complete
        CLEAR Decoder
        PUSH Decoder instance to Queue
ENDIF

IF Hypothesis not NULL
        RETURN Hypothesis
ENDIF

REPEAT
        Decoder = Decoder instance from Queue
        IF Decoder.LIVING-TIME Expires
                DELETE Decoder from Queue
        ENDIF
UNTIL Last element of Queue
```

D. *Evaluation and Feedback*

The transcription result is obtained when a particular recognition instance completes recognizing the given audio snippet. The transcription result is the word that has a greater confidence value in the dictionary of the speech recognition engine. Any mistakes in the pronunciation end up in incorrect result or a failure to recognize is likely to produce an error. This result is further evaluated against a predefined library of words.

Each result is compared in the order of its appearance in the articulated speech against the original set of words in the selected conversation. Any mismatch in the order of words can be identified and this is then highlighted in the display. A green highlighted word represents that which is correctly pronounced and appears in the right order. The wrongly pronounced word or the one that is not in the order of its appearance is highlighted in red. Whenever a word highlighted in red is encountered, the learner can continue that wrongly pronounced word correcting it and advancing the articulation. This gives the learner with a choice to correct the mistake and continue without the need of repeating the entire speech from the beginning.

V. CONCLUSION

Tablet education is highly motivational compared to traditional education methodologies. Offering the liberty to progress through the contents to meet the learning objectives at one's own pace makes learning more engaging and motivating. Such learner-centric methodology also requires appropriate evaluation and immediate feedback mandatory. Specific approach to language learning mentioned here, allows learners to rectify mistakes / errors and progress ahead instead of having to repeat from the very beginning every time should some mistakes occur. Proposed techniques such as Silence detection and Multiple Synchronized recognition enhances existing implementations by their ability to listen to longer articulated speech and recognize the words in real-time. While many speech recognition implementations has its recognition process proportional to the length of the articulated sentence, our method is able to outdo this limitation by recognizing words in a sentence almost instantly irrespective of the length of the articulated speech. Instant recognition with considerable accuracy facilitates immediate feedback possible, making ours a desirable language-learning tool. This is yet another addition to the suite of learning applications under the literacy initiative, Amrita Rural India Tablet-enhanced Education (RITE). Future work will explore the possibilities

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

of refining the dictionary, optimizing the language and acoustics model to facilitate recognition of Indian languages such as Hindi, Malayalam etc.

ACKNOWLEDGMENT

This work derives its inspiration and direction from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. We are grateful for the support of our colleagues at Amrita CREATE and the staff at Amrita University.

REFERENCES

[1] Nedungadi, P., Jayakumar, A., & Raman, R. (2015). "Low Cost Tablet Enhanced Pedagogy for Early Grade Reading: Indian Context". IEEE Region 10 Humanitarian Technology Conference. Chennai, pp. 35 - 39.

[2] Delmonte, R. (2002). A prosodic module for self-learning activities. In *Speech Prosody 2002, International Conference*.

[3] Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology*, *2*(1), 45-60.

[4] Uzun, L. (2012). The Internet and computer enhanced foreign language learning and intercultural communication. *World Journal on Educational Technology*,*4*(2), 99-112.

[5] Zhou, W., Zheng, J., Lu, Q., Chiu, T., You, X., & Ye, W. (2007, August). A computer assisted language learning system based on error trends grouping. In*Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007. International Conference on* (pp. 256-261). IEEE.

[6] Srikanth, R., & Salsman, L. B. J. (2012, December). Automatic Pronunciation Evaluation And Mispronunciation Detection Using CMUSphinx. In *24th International Conference on Computational Linguistics* (p. 61).

[7] Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., ... & Cesari, F. (2000). The SRI EduSpeakTM system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*, 123-128.

[8] Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996, October). Automatic text-independent pronunciation scoring of foreign language student speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 3, pp. 1457-1460). IEEE.

[9] Morton, H., Gunson, N., & Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction*, *2012*, 23.

[10] Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2), 95-108.

[11] Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication*, *51*(10), 832-844.

[12] Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *38*(1), 35-45.

[13] Nedungadi, P., & Raman, R. (2012). A new approach to personalization: integrating e-learning and m-learning. *Educational Technology Research and Development*, *60*(4), 659-6