ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization *Volume 4, Special Issue 11, September 2015*

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

Developing Telugu Speech Recognition System using Sphinx-4

[1]K.Venkataramana, [2]P.Jayaprakash, [3]P.S. Nagendra Babu

Research Scholar, Dept of CSE, Department of CSE, Dr.RR &Dr.SR Technological University, Chennai, Tamil Nadu, India

Assistant Professor, Dept of CSE, Department of CSE, Dr.RR &Dr.SR Technological University, Chennai, Tamil Nadu, India

Lecturer, Dept of CSE, Govt. Polytechnic College, Chennai, Tamil Nadu, India

ABSTRACT: Speech is the main communication medium in Human beings. Speech recognition has many applications. It can be used to automate many tasks that previously required hands-on human interaction. Many speech recognition systems have been proposed and developed. Sphinx4 is one such system developed in Java for recognizing English. In this paper how speech recognition is done is presented briefly and then sphinx4 architecture is explained. This paper aims at installing sphinx4 and testing it. Sphinx4 was used to develop applications for Telugu. This is done by identifying the language dependent areas of the sphinx4 and their interactions with other parts of the system and then modifying these areas so that it works for Telugu. Static Dictionary for Telugu words is developed rather than using online 'lmtool' of CMU. Sphinx4 was adapted to train and decode for recognizing isolated Telugu words. Continuing in a similar manner, a speech recognition system for Telugu can be developed.

KEYWORDS: Melfrequency cepstral coefficients,hidden markov models,lanuage model,n-grams

I. INTRODUCTION

Introduction To Speech Recognition

Speech is the primary mode of communication among humans. Our ability to communicate with machines and computers, through the keyboards, mice and other devices is some what slow in magnitude and also more cumbersome. So, to make this communication more user friendly, speech input is an essential component. Hence speech is the easiest way to interact with computer.Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, and having capability of correctly recognize what you are saying i.e., A System or software capable of recognizing spoken language. The machine or software may take the spoken language and translate it into text, or follow the spoken instructions to perform other functions. Simply it is the ability of a computer to recognize and understand human speech. In order to interface with computer using the speech mode, the system should be able to recognize the spoken data. Different speakers will articulate in differently due to Extrinsic and Intrinsic factors.

Extrinsic factors are those that were learned by the individual which was based on the primarily cultural factors such as environment where one grew up, the level of education attained and emotional factors such as the Speech. Types of Speech Recognition Depending upon the utterance of the different speakers, speech recognition classifies into different categories. These categories are based on the fact that one of the difficulties of speech recognition is the ability to determine when a speaker starts and finishes an utterance. Some of the categories are listed here:

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

Isolated Words:

Isolated words recognizes usually require each utterance to have a silence on both sides of the sample window. It does not mean that it accepts single words, but does require a single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances.

Connected Words:

Connected word systems are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

Continuous Speech:

Recognizers with continuous speech capabilities are one of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognition systems allow users to speak almost naturally, while the computer determines the content. It is basically computer dictation.

Spontaneous Speech:

There are many definitions for the spontaneous speech. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. A Speech recognition system with spontaneous speech ability should be able to handle of natural speech features such as words being run together.

Speaker Verification and Identification systems:

A speaker verification system must decide if a speaker is the person he/she claims to be. Such a system is otentially applicable to situations requiring control of access to information or restricted areas and to various kinds of automated credit transactions. A speaker identification system must decide which speaker among an ensemble of speakers produced a given speech utterance. Such systems have potential forensic applications.

II. ADAPTING TELUGU FOR SPHINX-4

3.1 Introduction

Sphinx-4 is an open source speech recognition system. It was created via a joint collaboration between the Sphinx group at Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs (MERL), and Hewlett Packard (HP), with contributions from the University of California at Santa Cruz (UCSC) and the Massachusetts Institute of Technology (MIT). Open source speech recognition systems are available, such as HTK, ISIP, and AVCSR and earlier versions of the Sphinx systems. The available systems are typically optimized for a single approach to speech system design. As a result, these systems intrinsically create barriers to future research that departs from the original purpose of the system.

The Sphinx-4 framework has been designed with a high degree of flexibility and modularity. Figure3.1 shows the overall architecture of the system. Each labeled element in Figure 3.1represents a module that can be easily replaced.

There are three primary modules in the Sphinx-4 framework: the FrontEnd, the Decoder, and the Linguist. The FrontEnd takes one or more input signals and parameterizes them into a sequence of Features. The Linguist translates any type of standard language model, along with pronunciation information from the Dictionary and structural information from one or more sets of AcousticModels, into a Search Graph.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

3.2 Sphinx-4 Architecture

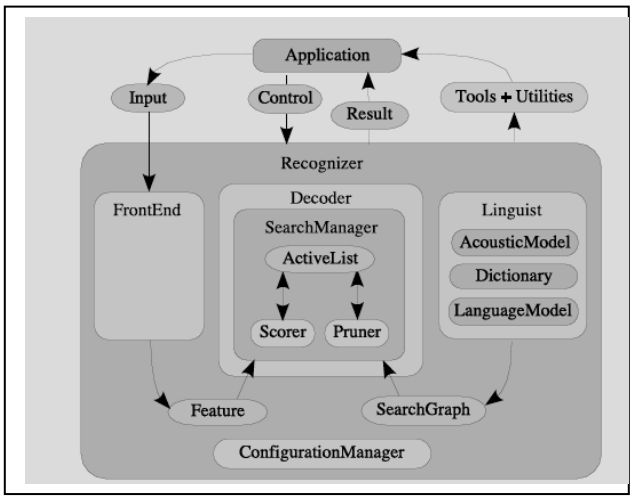


Fig1: Sphinx-4 Architecture. The communication between the blocks, as well as communication with an application

The SearchManager in the Decoder uses the Features from the Front End and the SearchGraph from the Linguist to perform the actual decoding, generating Results. At any time prior to or during the recognition process, the application can issue Controls to each of the modules, effectively becoming a partner in the recognition process.

IV. LINGUISTIC FEATURES OF TELUGU

4.1 Telugu Script Details

Telugu language script originated from the ancient Brahmi script. The basic units of the writing system are referred to as "Aksharas". The properties of Aksharas are as follows:

(1) An Akshara is an orthographic representation of a speech sound in Telugu language;

(2) Aksharas are syllabic in nature;

(3) The typical forms of Akshara are V, CV, CCV and CCCV, thus have a generalized form of C*V;

(4) An Akshara always ends with a vowel;

(5) White space is used as word boundary thus separating Aksharas present in two successive words;

(6) The scripts are written from left to right;

(7) Roman digits (0...9) are used as numerals. Some of the languages have their own numeric symbols which are rarely used;

(8) English Punctuations marks such as comma, full stops are mostly used in writing.

Convergence and divergence of Indian language scripts

India is a multi-lingual nation with 17 recognized official languages. These languages are: Assamese, Tamil, Malayalam, Gujarati, Telugu, Oriya, Urdu, Bengali, Sanskrit, Kashmiri, Sindhi, Punjabi, Konkani, Marathi, Manipuri, Kannada and Nepali. Except Urdu and English, all of the remaining official languages have a common phonetic base, i.e.,

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

they share a common set of speech sounds. While all of these languages share a common phonetic base, some of the languages such as Hindi, Marathi and Nepali also share a common script known as Devanagari. But languages such as Telugu, Kannada and Tamil have their own scripts. The property that makes these languages separate can be attributed to the Phonotactics in each of these languages rather than the scripts and speech sounds. Phonotactics is the permissible combinations of phones that can co-occur in a language.

4.1.1. Shape of an Akshara

The shape of an Akshara depends on its composition of consonants and the vowel, and sequence of the consonants. In defining the shape of an Akshara, one of the consonant symbols acts as pivotal symbol. Depending on the context, an Akshara can have a complex shape with other consonant and vowel symbols being placed on top, below, before, after or sometimes surrounding the pivotal symbol. Ideally the basic rendering unit for Indian language scripts should be Aksharas themselves. However, a language such as Telugu has around 15 vowels and 36 consonants. To render Aksharas as a whole unit, it requires 540 CV units, 19440 CCV units and 699840 CCCV units. It is reasonable to assume that not all combinations of consonant clusters are allowed, but even then nearly more than 10000 Aksharas are needed as rendering units. Due to this large number of units, an Akshara is rendered by concatenating the consonant and vowel symbols. The following are the symbols used to render Aksharas by a Unicode [9] rendering engine.

Consonant symbol

A consonant symbol in an Indian language represents a single consonant sound and also an inherent vowel (short /a/). Akshara is syllabic so each consonant symbol represents a consonant and the inherent vowel.

Half forms of the consonant

Aksharas of the type CCV or CCCV, have more than one consonant. In these cases, the initial consonant(s) have half-form shape. These half-forms do not have an inherent vowel.

Vowel symbols (independent vowels)

A vowel symbol represents a vowel sound and is used to render a syllable of type V which has no consonants before or after it. These vowel symbols are also referred to as independent vowels.

Viraam

Sometimes it is necessary to write consonants without inherent vowels. To remove the inherent vowel from a consonant, a symbol called Viraam is used. The symbol may be an oblique stroke under the consonant symbol, or can change the shape of the consonant if it is on top of the consonant.

4.1.2. Unicode: A Universal Character Set

Computers can only interpret bits and bytes, and hence the representation of a script should be defined in terms of bits and bytes. ASCII—American Standard Code for Information Interchange is an 8-bit code to represent the English character set. Similarly there is Indian Standard Code for Information Interchange (ISCII) that defines an 8-bit character code for Indian language scripts. As these codes overlap, computers using ASCII character set cannot interpret ISCII as a code for Indian language scripts. With an 8-bit code, only 256 unique characters can be defined. To allow computers to represent any character in any language, the international standard ISO 10646 defines the Universal Character Set (UCS). UCS contains the characters to practically represent all known languages in the world. ISO 10646 originally defined a 32-bit character set. Each character is assigned a 32 bit code. However, these codes vary only in the least-significant 16 bits. ISO 10646 and Unicode though started as two projects finally merged their character set around 1991 so that both are now compatible with each other. In addition to the character set, Unicode standard specifies recommendation for rendering of the scripts, handling of bi-directional texts that mix for instance Latin (left to right writing system) and Hebrew (right to left writing system), algorithms for storage and manipulation of Unicode strings (Alan, 2005; The Unicode Consortium, 2003).

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

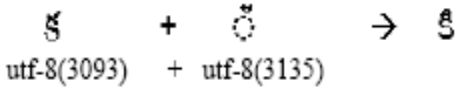
UTF-8 and UTF-16

It has to be noted that Unicode is a table of codes that assigns integer numbers to characters (Markus, 2005). One still has to define its implementation or encoding in the computers. A straightforward encoding of these integers is to store the Unicode text as sequences of 2 byte sequences. This encoding is referred to as UTF-16. An ASCII file can be transformed into a UTF-16 file by simply inserting a 0x00 byte in front of every ASCII byte.

Representation of Indian language scripts

Having known the syllabic nature of Indian language scripts, it is easy to understand the notation followed by the Unicode to represent Indian language characters. The following are the principles used by Unicode to represent the Indian language characters:

(1) A Unicode is assigned to each consonant symbol, i.e., a consonant sound along with the inherent vowel.
(2) Each independent vowel is represented by a Unicode.
(3) Each Maatra is also represented by a Unicode.
(4) Viraam has a Unicode number too.



The Unicode 3093 represents the full consonant symbol, and the Unicode 3135 represents the Maatra. The concatenation of a Maatra to the consonant symbol changes its shape to get the required CV unit. Any desired CV unit can thus be produced by a simple concatenation of UTF-8 representation of the consonant symbol and the Maatra

Rendering a CCV unit

This section defines the complete set of phones used in Telugu speech. It also includes feature definitions as follows:

1. **Vowel or Consonant:**

In phonetics, a **vowel** is a sound in spoken language, such as Telugu అ or ఓ, pronounced with an open vocal tract so that there is no build-up of air pressure at any point above the glottis. This contrasts with **consonant**, such as Telugu క, where there is a constriction or closure at some point along the vocal tract.

2. **Vowel Length**: Duration of a Vowel sound, normally it is either short or long.

   a) **Short:** Where the duration of a vowel is short like, అ, ఇ, ఉ.

   b) **Long:** Where the duration of vowel is long like, ఆ , ఈ , ఊ

   c) **Dipthong:** Some Vowels are characterized by a movement from one vocal tract position to another. Ex: ఏ ఐ ఓ ఔ

   d) **Schwa:** Neutral Vowel, in that its position is neither front nor back, high nor low, rounded nor un-rounded. Ex: ఎ

3. **Vowel height:** It is the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw.

   a) high: ఇ ఈ ఊ

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

b) mid: ఉ ఎ ఏ ఒ ఓ

c) low: అ ఆ ఐ ఔ

4. **Vowel frontness:** The position of the raised part of the tongue with in relation to the front of the mouth defines either front or back.

a) **front:** ఇ ఈ ఏ

b) **mid:** ఎ ఐ ఒ ఓ ఔ

c) **back:** అ ఆ ఉ ఊ

5. **Lip Rounding:** This effect is caused when the lips are protected from their normal position. For instance, ఉ ఊ ఓ are rounded and remaining is un-rounded.

d) Nasal: Where there is complete occlusion of the oral cavity, and the air passes instead through the noise. The shape and position of the tongue determine the resonant cavity.

Ex: న మ

e) Liquid: These are more like normal consonants, but still have many unusual properties like smooth or freely moving of the sound.

Ex: ల ళ

7. Place of Articulation: It is the point of contact where an obstruction occurs in the vocal tract between an active articulation (moving, like tongue) and passive articulation (stationary, like roof of the mouth).

The figure 4.1 clearly shows the different place of articulations:

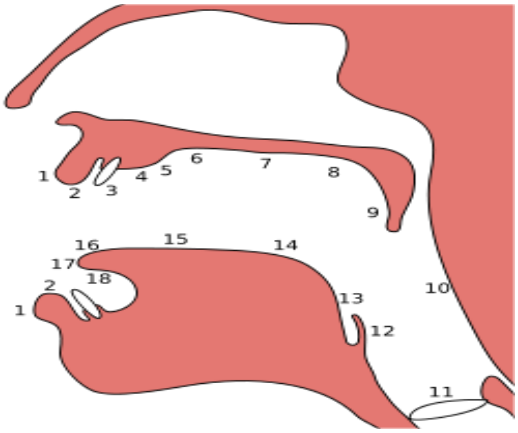


Fig 4.1 Places of articulation (passive & active): 1. Exo-labial, 2. Endo-labial, 3. Dental, 4. Alveolar, 5. Post-alveolar, 6. Pre-palatal, 7. Palatal, 8. Velar, 9. Uvular, 10. Pharyngeal, 11. Glottal, 12. Epiglottal, 13. Radical, 14. Postero-dorsal, 15. Antero-dorsal, 16. Laminal, 17. Apical, 18. Sub-apical

a) Labial: The active lower lip may contact either a passive upper lip

Ex: ప, ఫ, బ, భ, మ

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

b) Alveolar: Between the active front of the tongue and the ridge behind the gums (the alveolus)

Ex: త, ద

c) Palatal: Between the active middle of the tongue and the passive hard palate.

Ex: చ, న

d) Labio-dental: The active lower lip may contact either a passive upper teeth.

Ex: వ

e) Dental: Between the active front of the tongue and the passive upper teeth.

Ex: థ, ధ

f) Velar: Between the active back of the tongue and the passive soft palate

Ex: క, గ

g) Glottal: Between the larynx and vocal cords.

Ex: హ

The following fig 4.2 clearly shows the all the above types.

8. Consonant voicing: A voiced sound is one in which the vocal cords are vibrate, and a voiceless sound is one in which they do not.

Ex: Voiced: థ ధ న ఫ భ మ య ర ల

unvoiced: క చ త ప స

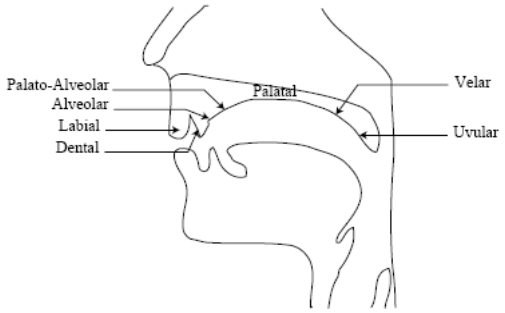


Fig 4.2 Telugu phonemes place of articulation with respect to building Telugu speech recognition system.

V. CONCLUSION

Isolated word speech recognition system for Telugu based on Hidden Markov Models (HMM) and Mel Frequency Cepstral Coefficient features (MFCC) has been developed using Sphinx4. Two hundred and fifty Telugu words recorded 3 times by

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization *Volume 4, Special Issue 11, September 2015*

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India.

20 speakers were used for training. Implemented Telugu static dictionary and phone list for training. Implemented Telugu JSGF as grammar file for Sphinx4 decoder. Then 10 different speakers spoken some list of words for performance evaluation. An average of 81% accuracy was achieved.

VI. FEATURE DIRECTIONS

In the future performance of the system should be tested on large amount of speech data. For developing large vocabulary speech recognition system Telugu static dictionary must be enhanced. By implementing N-grams it is possible to develop Telugu continuous speech recognition system using sphinx-4. Domain specific applications like mobile phone customer and caller tune services and railway reservation systems can be implemented using Telugu speech recognition system. Possible to implement some information retrieval systems for formers in Telugu language. Possible to implement some Telugu word document editors for text processing.

REFERENCES

[1].W.A.Lea, " Trends in Speech Recognition", Prentice Hall, 1980.
[2]. http://www-2.cs.cmu.edu/~robust/Tutorial
[3]. Modelling Word Duration for Better Speech Recognition by Venkata Ramana Rao Gadde SRI International Menlo Park, USA
[4]. http://www.cis.hut.fi/jpylkkon/pylkonen04norsig.pdf
[5]. http://www.utdallas.edu/~loizou/cimplants/hillen.pdf
[6]. http://www.mmk.e-technik.tu-muenchen.de/pb/ps/00pfa1.pdf
[7].http://research.microsoft.com/srg/papers/1999-richardsoneurospeech.pdf