Challenges in developing Large Vocabulary Speech Recognition System for Telugu

Dr. G. Bharadwaja Kumar

AU-KBC Research Centre,

Chennai.

Contents

- Basics of Speech recognition
- Sphinx-4 speech recognition system
- Experiments and Results of Telugu Speech Recognition System
- Analysis and Discussion

Definition

- **Speech Recognition** is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words.

Speech recognition systems can be characterized by many parameters

Parameters	Range
Speaking Mode	Isolated words to Continuous Speech
Speaking style	Read Speech to Spontaneous Speech
Enrollment	Speaker dependent to Speaker independent
Vocabulary	Small (<20 words) to large (> 60,000 words)
Language model	Finite State to Context Sensitive
Perplexity	Small (<10) to large (>100)
Signal to Noise Ratio (SNR)	High (>30dB) to low (<10dB)
Transducer	Voice cancelling microphone to Telephone

Typical parameters used to characterize the capability of speech recognition systems

Speech Recognition Modalities

- **Isolated Word Recognition**
  - Each word is assumed to be surrounded by silence
  - "this…is…isolated…word…recognition"
- **Connected-Word Recognition**
  - Word sequences constrained by a fixed grammar (e.g., telephone numbers)
- **Word Spotting**
  - Detect word in presence of surrounding words
  - "this is word spotting"
- **Continuous Speech Recognition**
  - Fluent, uninterrupted speech

Why Continuous speech recognition is difficult?

- Word boundaries are unclear
- Continuous speech less clearly articulated
- Co-articulation and phonetic context impacts speech both within words and across word boundaries

Source-Channel model

- If **A** represents the acoustic feature sequence extracted from a text sample, the speech recognition system should yield the optimal word sequence which matches **A** best .

$$\hat{W} = \underset{W}{\operatorname{argmax}}\ P(W|A)$$

- Using Baye's rule, we can rewrite as

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

- Here, **P(A|W)** is the likelihood of feature sequence **A** given the acoustic model of word sequence **W**.

- The acoustic model consists of the speech signal features to be used for **O**, and a pattern matching technique to compare these features against a set of predetermined patterns of these features for a given word or phone.

- **P(W)** is the probability of word sequence which is computed from the language model. The language model of an ASR system predicts the likelihood of a given word sequence appearing in a language.

Goals of Feature Extraction

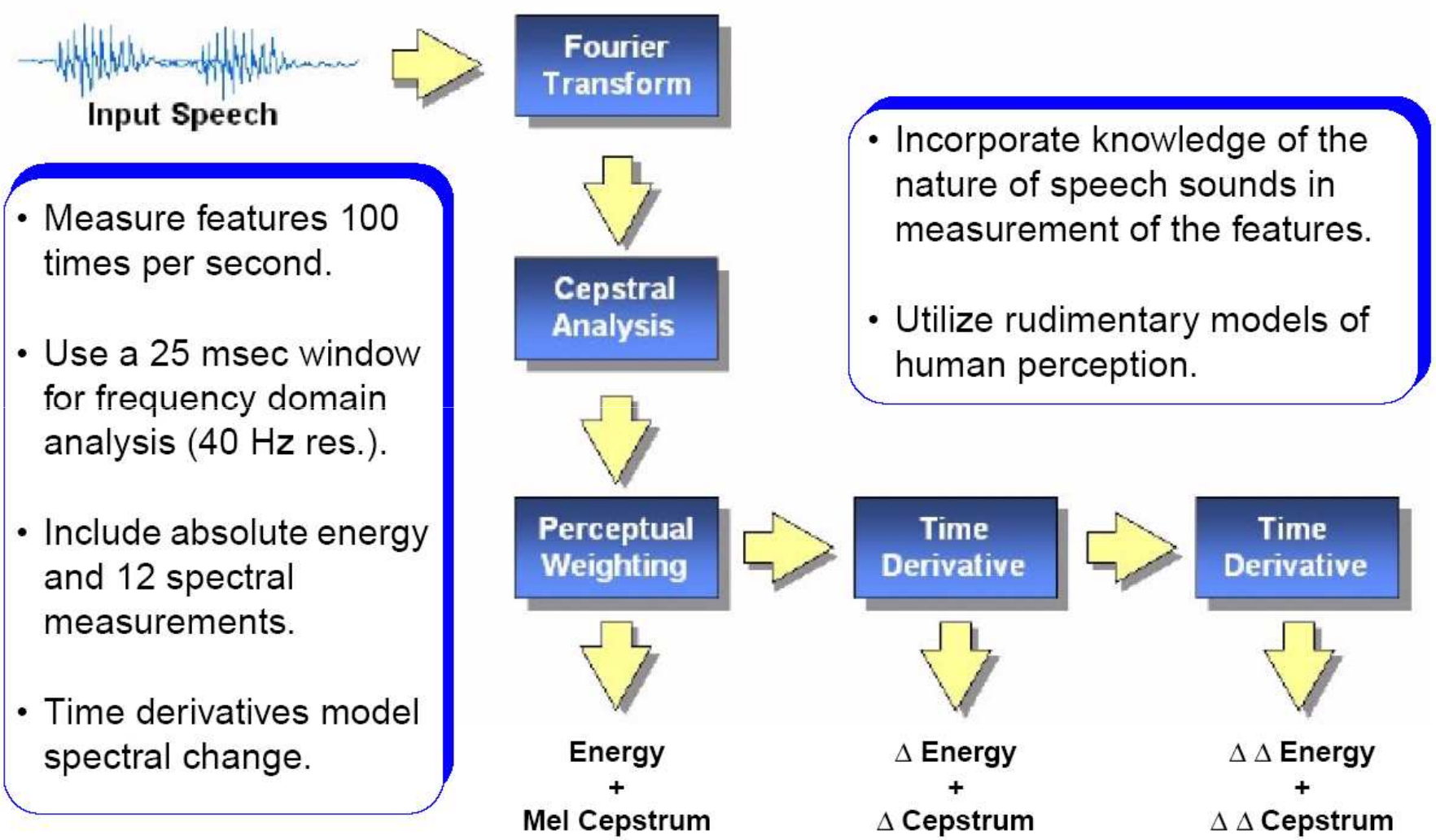
- Compactness
- Discrimination Power
- Low Computation Complexity
- Reliable
- Robust

- A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Cepstral coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC), and Perceptual linear prediction (PLP).

- MFCC is perhaps the best known and most popular because MFCC's are based on the known variation of the human ear's critical bandwidths with frequency.

- Advantage : Less coefficients, coefficients are uncorrelated

Feature Extraction in Speech Recognition



- Measure features 100 times per second.
- Use a 25 msec window for frequency domain analysis (40 Hz res.).
- Include absolute energy and 12 spectral measurements.
- Time derivatives model spectral change.

- Incorporate knowledge of the nature of speech sounds in measurement of the features.
- Utilize rudimentary models of human perception.

Acoustic Models for Speech Recognition

HMMs for Speech

- Each state can be associated with
  - sub-phoneme
  - phoneme
  - sub-word

- Usually, sub-phonemes or sub-words are used, to account for co-articulation (spectral dynamics).

- One HMM corresponds to one phoneme or word

- For each HMM, determine the most likely state sequence that results in the observed speech.

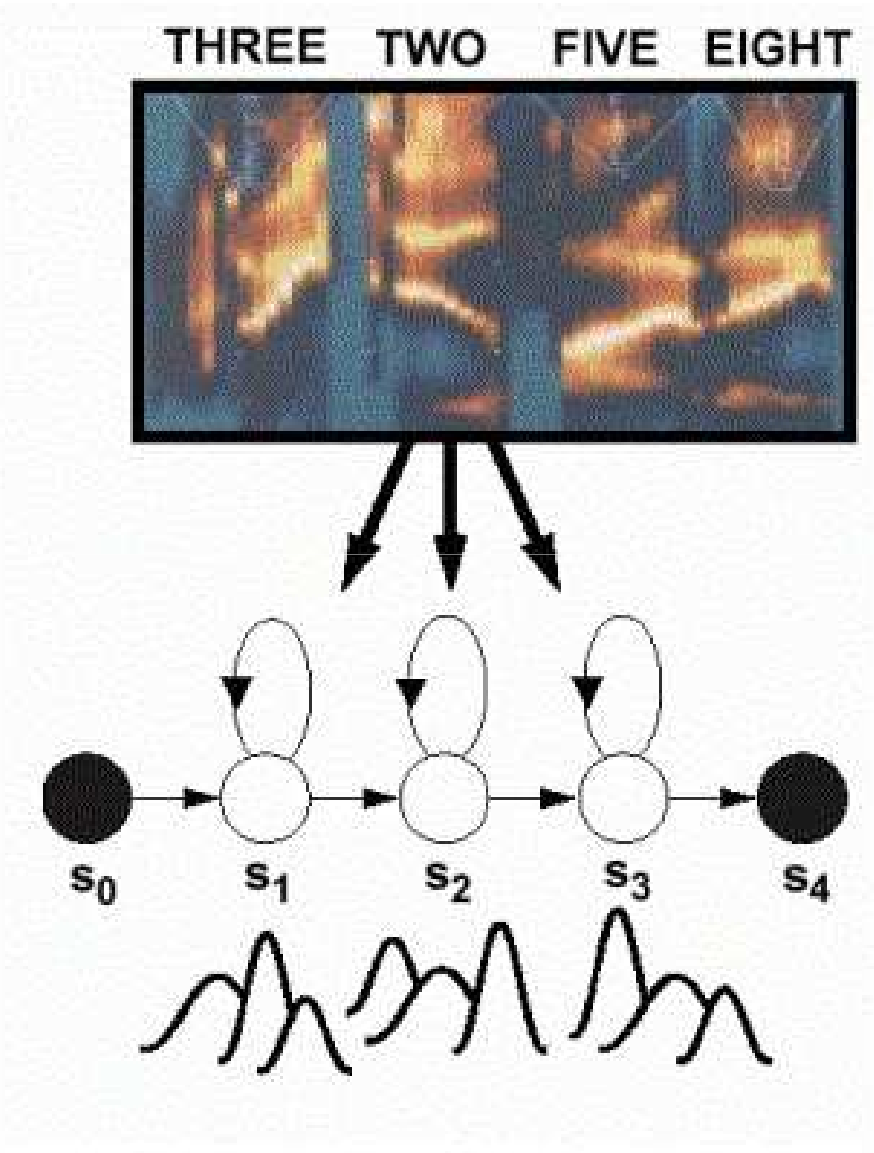
- Choose HMM with best match to observed speech.

Training HMMs for Continuous Speech

- Use only orthograph transcription of sentence
  - no need for segmented/labelled data
- Concatenate phone models to give word model
- Concatenate word models to give sentence model
- Train entire sentence model on entire spoken sentence

Acoustic Modeling: Hidden Markov Models

- Acoustic models encode the temporal evolution of the features (spectrum).
- Gaussian mixture distributions are used to account for variations in speaker, accent, and pronunciation.
- Phonetic model topologies are simple left-to-right structures.
- Skip states (time-warping) and multiple paths (alternate pronunciations) are also common features of models.
- Sharing model parameters is a common strategy to reduce complexity.



Whole-Word HMMs

- **Assign a number of HMM states to model a word as a whole.**
- **Passes the test?**
  - Accurate – Yes, if you have enough data and your environment consists of a small vocabulary. No, if you are trying to model context changes between words.
  - Compact – No, need too many states as vocabulary increases. Probably not enough training data to model *every* word. What about infrequent words???
  - General – No, can't build new words using this representation.

Context-Independent Phoneme HMMs

- **Context-independent models consist of a single *M*-state HMM (e.g., *M*=3), one for each phoneme unit**
- **Also referred to as "monophone" models**
- **Passes the test?**
  - Accurate – No, does not accurately model coarticulation
  - Compact – Yes, *M* states x *N* phonemes leads to only a few parameters which need to be estimated.
  - General – Yes, you can construct new words by stringing together the units.

Context-Dependent Triphone HMMs

- **Context-dependent models which consist of a single 3-state HMM, one for each phoneme unit modeled with the immediate left-and-right phonetic context**

- **Passes the test?**
  - Accurate – Yes, takes coarticulation into account
  - Compact – Yes, Trainable – No: For N phonemes, there exists NxNxN triphone models. Too many parameters to estimate!
  - General – Yes, you can construct new words by stringing together the units.

Language Models for Speech Recognition

Language Model (LM)

- A collection of prior knowledge about a language which is independent of an utterance to be recognized.
- Represents previous knowledge about language and the expectations at utterances.
- Can be expressed in terms of which words or word sequences are possible or how frequently they occur.

- Need for a language model in a speech recognizer arises from the variability of the speech signal, missing word boundaries and homographs.

- Needed to resolve the ambiguities which the acoustic model is not able to handle.

- **Statistical language models:** If the model is based on counting events in a large text corpus, for example, how frequent a certain word or word sequence occurs, the model is called to be a statistical language model.

- **Knowledge based models:** If the knowledge comes from a human expert the model is called knowledge-based language model. (Rule-Based Language models

N-Gram models

- Given a text $w_1^T = w_1, \ldots w_t \ldots w_T$, we can compute its probability by

$$\Pr(w_1^T) = \Pr(w_1) \prod_{t=2}^{T} \Pr(w_t \mid h_t)$$

where $h_t = w_1 \ldots w_{t-1}$ indicates history of the word w_t

- $\Pr(w_t \mid h_t)$ is difficult to estimate as the history of h_t grows

- Hence we restrict the history to bigrams or trigrams

N-gram Language Model

- Widely used statistical language models
- Probabilities are estimated from a corpus of training data (text data).
- Once model is known, new sentences can be randomly generated by the model!
- Syntax roughly encoded by model, but ungrammatical and semantically “strange” sentences can be produced

- larger n: more information about the context of the specific instance (greater discrimination)

- smaller n: more instances in training data, better statistical estimates (more reliability)

Language Models: Class-Based LM

- Category-Based, Class-Based, or Clustering LM improve the number of counts (and therefore the robustness) by grouping words into different classes.

- In one case, all relevant words belonging to one category can be clustered into one class. In the language model, the class is treated as a "normal" word.

- For example, P("January" | w_1, w_2) is considered comparable to P("February" | w_1, w_2) or any other month. Rather than having separate probability estimates for w_1, w_2 followed by each month (some months may not occur at all in the training data), collapse all months into the single class "month_class", and compute P("month_class" | w_1, w_2)

Language Models: Class-Based LM

- In another case, *all* words are assigned to a class (e.g. semantic category or part of speech such as noun, verb, etc.). Then, if C_i is the class for word w_i, the trigram language model is computed using one of:

$$P(w_3 \mid w_1, w_2) = P(w_3 \mid C_3) \cdot P(C_3 \mid w_1, w_2) \quad \textbf{[1]}$$

$$P(w_3 \mid w_1, w_2) = P(w_3 \mid C_3) \cdot P(C_3 \mid w_1, C_2) \quad \textbf{[2]}$$

$$P(w_3 \mid w_1, w_2) = P(w_3 \mid C_3) \cdot P(C_3 \mid C_1, C_2) \quad \textbf{[3]}$$

$$P(w_3 \mid w_1, w_2) = P(w_3 \mid C_1, C_2) \quad \textbf{[4]}$$

- Improvement in performance depends on how clustering is done (manually or automatically, semantic categories or part-of-speech categories) and how trigram probabilities are computed (using one of **[1]** through **[4]** or some other formula).

Factored Language Models

Generalised language models in which
word and word related features (factors)
are bundled together

Nearly everything!!
Word themselves, stems, Part Of Speech tags, relative position in the sentence, morphological classes, DAs, ….

Goal: factorise the joint probability associated to a sentence in terms of factor-related conditional probabilities:

$$p(w_{1:T}) = \prod_t p(w_t, f_t^1, \ldots, f_t^k \mid w_{t-1}, f_{t-1}^1, \ldots, f_{t-1}^k, w_{t-2}, f_{t-2}^1, \ldots, f_{t-2}^k, \ldots, f_{t-n}^{0:k})$$

Smoothing

- Smoothing addresses data sparsity problem
- There is rarely enough data to estimate parameters of the language models
- Smoothing gives a way to combine less specific, more accurate information with more specific but noisy data.

- Want some way to combine trigram probabilities, bigram probabilities, unigram probabilities
- Use trigram/bigram for accurate modeling of context
  - Use unigram to get best guess you can when data is sparse.
- Lots of different techniques
  - Simple interpolation
  - Absolute Discounting
  - Katz Smoothing (Good-Turing)
  - Interpolated Kneser-Ney Smoothing

Perplexity

- One popular measure of the difficulty of the task, combining the vocabulary size and the language model, is *perplexity, loosely defined as the geometric mean* of the number of words that can follow a word after the language model has been applied.

- Perplexity measures both the quality of the language model (better language models yield lower PP values on the same data) and the difficulty of the task (harder tasks yield larger PP values).

- Reduction in perplexity does not always correspond to reduction in word error rate, but *PP* is simple and convenient measure.

Perplexity:
Is lower better?

- Remarkable fact: the true model for data has the lowest possible perplexity
- Lower the perplexity, the closer we are to true model.
- Typically, perplexity correlates well with speech recognition word error rate
  - Correlates better when both models evaluated on same data
  - Doesn’t correlate well when training data changes

Evaluation Metric For Speech recognition System

Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N}$$

where

- S is the number of substitutions,
- D is the number of the deletions,
- I is the number of the insertions,
- N is the number of words in the reference.

HTK

- HTK is the “Hidden Markov Model Toolkit” developed by the Cambridge University Engineering Department (CUED).

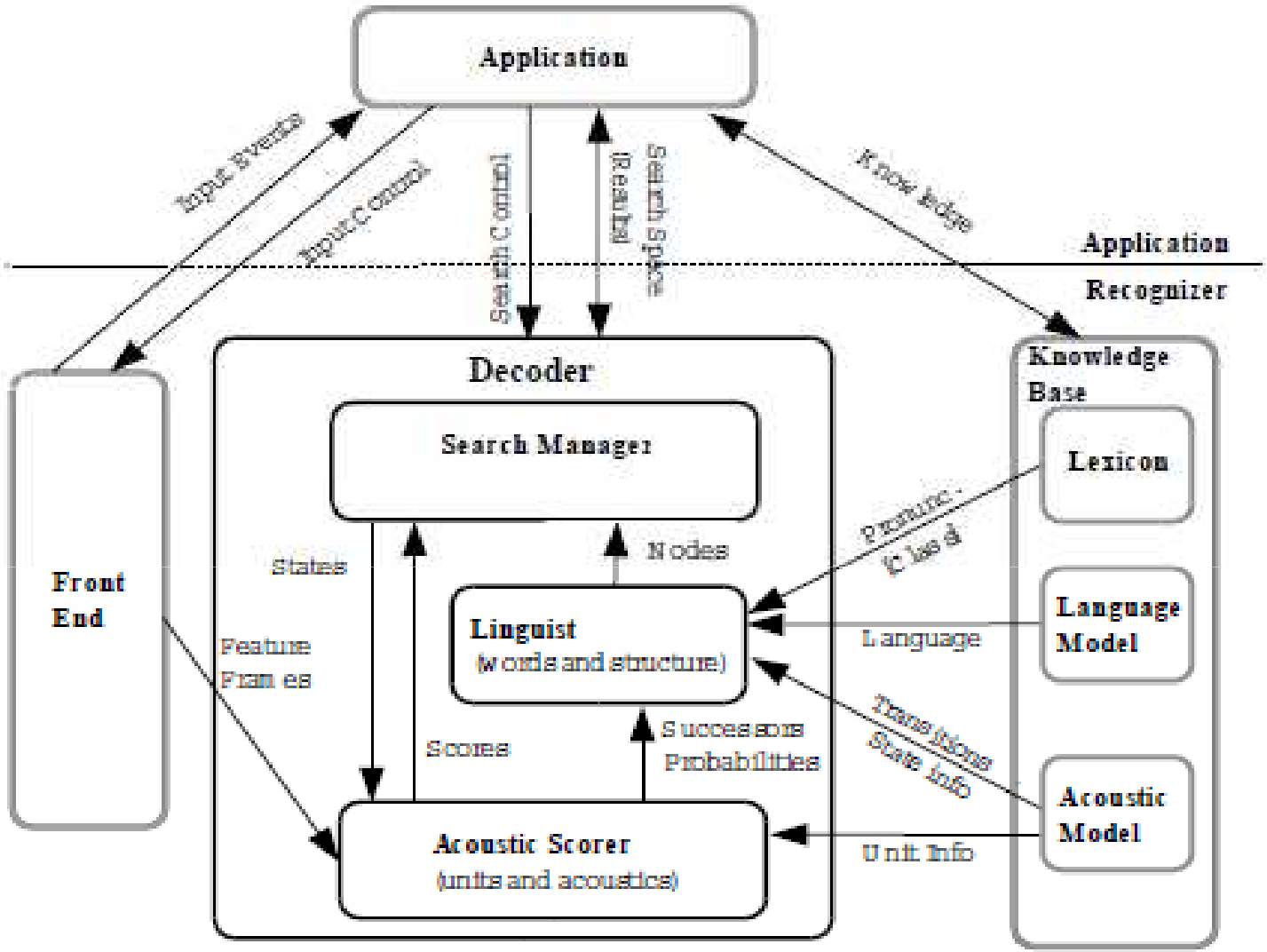
- HTK is primarily used for **speech recognition research**
- http://htk.eng.cam.ac.uk/

SPHINX-4

- The Sphinx-4 speech recognition system has been jointly developed by Carnegie Mellon University, Sun Microsystems Laboratories, and Mitsubishi Electric Research Laboratories (MERL).

- Sphinx-4 has been developed in the Java programming language.

- http://cmusphinx.sourceforge.net/html/cmusphinx.php



Sphinx-4 system architecture.

- It is highly modular and flexible, supporting all types of HMM-based acoustic models, all standard types of language models, and multiple search strategies.

- *Features:* MFCC, PLP
  - **Feasible to add our own features**

- One can, for instance, change the language model from a statistical N-gram language model to a context free grammar (CFG) or a stochastic CFG by modifying only one component of the system, namely the *linguist*

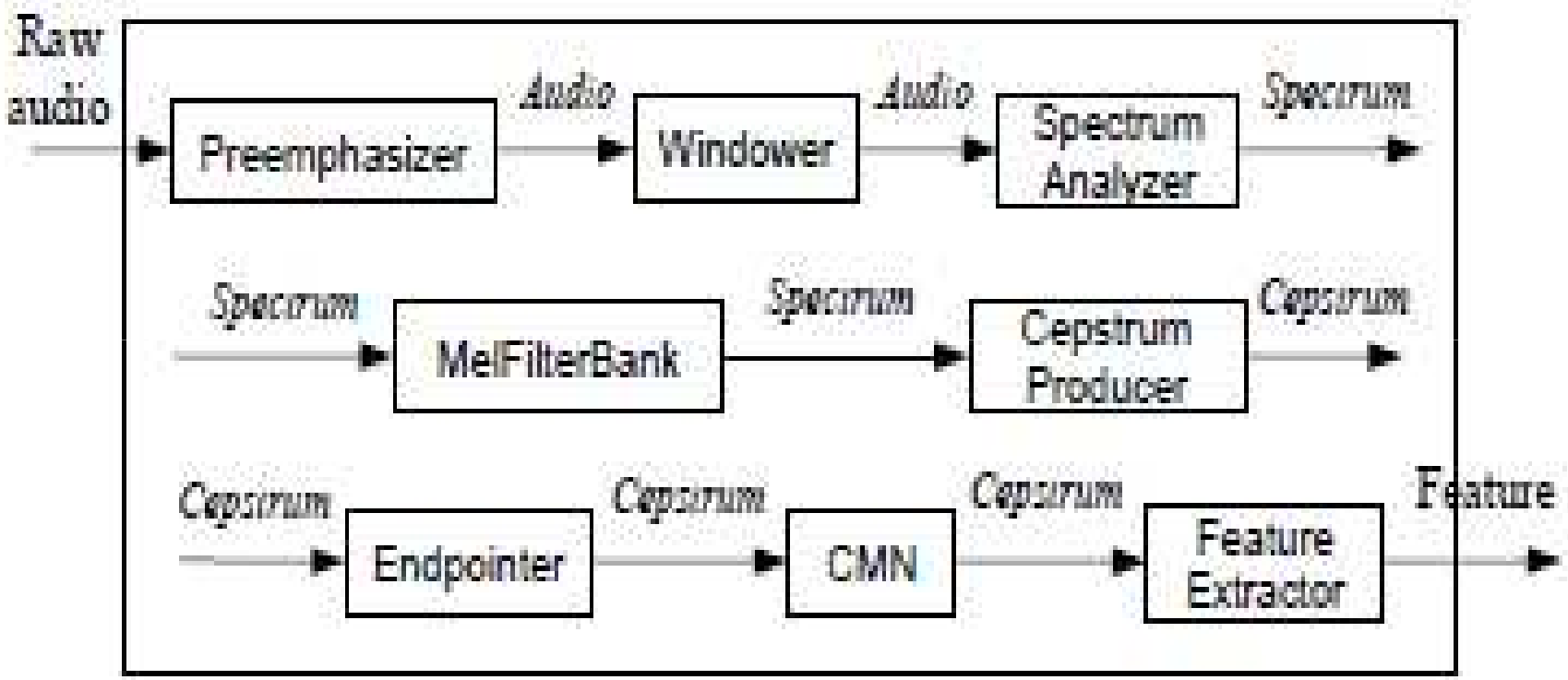
- It is possible to run the system using continuous, semi-continuous or discrete state output distributions by appropriate modification of the *acoustic scorer*

- The system permits the use of any level of context in the definition of the basic sound units

- Information from multiple information streams can be incorporated and combined at any level, *i.e., state, phoneme, word or grammar*

- The search module can also switch between depth-first and breadth-first search strategies

Sphinx-4 front end



Front end

- The FrontEnd takes one or more input signals and parameterizes them into a sequence of Features.

Linguist

- 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 SearchGraph.
- The AcousticModel module provides a mapping between a unit of speech and an HMM that can be scored against incoming features provided by the FrontEnd.

- Linguist breaks each word in the active vocabulary into a sequence of context-dependent sub-word units. The
- Linguist then passes the units and their contexts to the AcousticModel, retrieving the HMM graphs associated with those units.
- It then uses these HMM graphs in conjunction with the LanguageModel to construct the SearchGraph.

Decoder

- The SearchManager in the Decoder uses the Features from the FrontEnd and the SearchGraph from the Linguist to perform the actual decoding, generating Results.

Specifications of Training Data

- Data collected in general lab environment recorded with ordinary microphone over computer
- Consists of 720 sentences from newspapers which are phonetically rich
- 20 male speakers and 8 female speakers
- Few speakers spoke all 720 sentences and rest each speaker has spoken a minimum of 180 sentences each.
- The speakers age is mostly in between 20-35.
- The total training data comprises to nearly 18 hours of speech data.

Specifications of Test Data

- 100 sentences from the test corpus that comprises to 18,000 vocabulary.
- 4 trained voices (2 male and 2 female) each spoken 25 sentences
- 10 untrained voices (8 male and 2 female) where each speaker has spoken ten sentences.

Performance of our Telugu speech recognition system (16KHz, 16bit)

Details	Trained Voices	Untrained voices
No. of Sentences in test data	100	100
No. of words in test data	998	998
Error Types	Substitutions: 70 Deletions: 18 Insertions: 7	Substitutions: 144 Deletions: 19 Insertions: 40
WER(word error rate)	9.52%	20.34%
SER (sentence error rate)	31%	51%

Telephonic Speech

- Data has been collected over Landline using analog voice recorder
- Only three speakers data available for training
- Tested for one trained voice

Performance of our Telugu speech recognition system (8KHz, 16bit)

Details	Trained Voices
No. of Sentences in test data	50
No. of words in test data	406
Error Types	Substitutions: 104 Deletions: 8 Insertions: 3
WER(word error rate)	28.33%
SER (sentence error rate)	66%

Analysis of Results

- Numeric numbers and their written word forms both are the present in the corpus

REF: ఇవిగాక ఫ్యాప్సియా రూ 52 లక్షల వస్తుసామగ్రి ఆపాలో ఇరవైవేడు లక్షల 470 ఆ రూపాయిల విలువైన ఔషధాలను అందజేసినట్లు తెలిపింది

HYP: ఇవిగాక ఫ్యాప్సియా రూ 52 లక్షల వస్తుసామగ్రి ఆపాలో 2500436 రూపాయిల విలువైన ఔషధాలను అందజేసినట్లు తెలిపింది

ALIGN_REF: ఇవిగాక ఫ్యాప్సియా రూ 52 లక్షల వస్తుసామగ్రి ఆపాలో ఇరవైవేడు లక్షల 470 ఆ రూపాయిల విలువైన ఔషధాలను అందజేసినట్లు తెలిపింది

ALIGN_HYP: ఇవిగాక ఫ్యాప్సియా రూ 52 లక్షల వస్తుసామగ్రి ఆపాలో 2500436 రూపాయిల విలువైన ఔషధాలను అందజేసినట్లు తెలిపింది ******* ********** *********

Accuracy: %43.8 Errors: 9 (Sub: 6 Ins: 0 Del: 3)
Words: 16 Matches: 7 WER: %56.2
Sentences: 1 Matches: 0 SentenceAcc: %.0

REF: మంత్రి పి.సుదర్శన్‌రెడ్డి ఆధ్వర్యంలో వివిధ సంస్థల మంది గురుపు 2967924 సీఎం సహాయ నిధికి విరాళాలుగా అందాయి

HYP: మంత్రి పి.సుదర్శన్‌రెడ్డి ఆధ్వర్యంలో వివిధ సంస్థల నుండి రూ 2967924 సీఎం సహాయ నిధికి విరాళాలుగా అందాయి

ALIGN_REF: మంత్రి పి.సుదర్శన్‌రెడ్డి ఆధ్వర్యంలో వివిధ సంస్థల మంది గురుపు 2967924 సీఎం సహాయ నిధికి విరాళాలుగా అందాయి

ALIGN_HYP: మంత్రి పి.సుదర్శన్‌రెడ్డి ఆధ్వర్యంలో వివిధ సంస్థల నుండి రూ 2967924 సీఎం సహాయ నిధికి విరాళాలుగా అందాయి

```
Accuracy: %84.6    Errors: 2  (Sub: 2  Ins: 0  Del: 0)
Words: 13   Matches: 11    WER: %15.4
Sentences: 1   Matches: 0   SentenceAcc: %.0
```

- One more reason is that untrained voices contain non-native speakers of Telugu who does not have any idea about Telugu language

  (transliteration has been used for recording of non-native speakers)
- Need to extend our phone set to deal with transliterated words which has not been studied carefully

REF: మార్కండేయుడు తనకు అల్పాయుష్షు ఉందని విన్నా ఏమాత్రం కలత చంద్ర కుండ తల్లిదండ్రుల దగ్గర సెలవు తీసుకుని పరమశివుడిని మెప్పించి చిరంజీవి కావాలని తపస్సు చేసుకోవడానికి వెళ్లాడు ఆరో

HYP: మార్కండేయుడు తనకు అల్పాయుష్షు ఉందని విన్నా ఏ మాత్రం కలత చెందకుండా తల్లిదండ్రుల దగ్గర సెలవు తీసుకొని పరమశివుడిని మెప్పించి చిరంజీవి కావాలని తపస్సు చేసుకోడానికి వెళ్ళాడు

ALIGN_REF: మార్కండేయుడు తనకు అల్పాయుష్షు ఉందని విన్నా ఏమాత్రం కలత చంద్ర కుండ తల్లిదండ్రుల దగ్గర సెలవు తీసుకుని పరమశివుడిని మెప్పించి చిరంజీవి కావాలని తపస్సు చేసుకోవడానికి వెళ్లాడు ఆరో

ALIGN_HYP: మార్కండేయుడు తనకు అల్పాయుష్షు ఉందని విన్నా ఏ మాత్రం కలత చెందకుండా తల్లిదండ్రుల దగ్గర సెలవు తీసుకొని పరమశివుడిని మెప్పించి చిరంజీవి కావాలని తపస్సు చేసుకోడానికి వెళ్ళాడు ***

```
   Accuracy: %61.9    Errors: 8  (Sub: 7  Ins: 0  Del: 1)
   Words: 21   Matches: 13    WER: %38.1
   Sentences: 1   Matches: 0   SentenceAcc: %.0
Creating new instance of LogMath while another instance is already present
```

REF: తన మనసు శ్రీమహావిష్ణువు సేవమీద లగ్నమై ఉండేలా వరం ఇవ్వమని పార్వతీ పరమేశ్వరులను మార్కండేయుడు వేడుకున్నాడు ఈ రూ

HYP: తన మనస్సు శ్రీమహావిష్ణువు సేవ మీద లగ్నమై ఉండేలా వరం ఇవ్వమని పార్వతీపరమేశ్వరులను మార్కండేయుడు వేడుకొన్నాడు

ALIGN_REF: తన ****** మనసు శ్రీమహావిష్ణువు సేవమీద లగ్నమై ఉండేలా వరం ఇవ్వమని పార్వతీ పరమేశ్వరులను మార్కండేయుడు వేడుకున్నాడు ఈ రూ

ALIGN_HYP: తన మనస్సు శ్రీమహావిష్ణువు సేవ మీద లగ్నమై ఉండేలా వరం ఇవ్వమని ******* పార్వతీపరమేశ్వరులను మార్కండేయుడు వేడుకొన్నాడు * **

```
Accuracy: %42.9    Errors: 9  (Sub: 5  Ins: 1  Del: 3)
Words: 14   Matches: 6    WER: %64.3
Sentences: 1   Matches: 0   SentenceAcc: %.0
```

REF: పూర్వకాలంలో విజ్ఞతకు ఇక అగ్నికి ఈ జ్ఞానానికి ప్రాధాన్యమిచ్చేవారు

HYP: పూర్వకాలంలో విజ్ఞతకు వివేకానికి జ్ఞానానికి ప్రాధాన్యమిచ్చేవారు

ALIGN_REF: పూర్వకాలంలో విజ్ఞతకు ఇక అగ్నికి ఈ జ్ఞానానికి ప్రాధాన్యమిచ్చేవారు

ALIGN_HYP: పూర్వకాలంలో విజ్ఞతకు వివేకానికి జ్ఞానానికి ప్రాధాన్యమిచ్చేవారు ********** **********

```
Accuracy: %28.6    Errors: 5  (Sub: 3  Ins: 0  Del: 2)
Words: 7   Matches: 2    WER: %71.4
Sentences: 1   Matches: 0   SentenceAcc: %.0
```

Difficulties in collecting data

- Many speakers have not shown variation between aspirated and un-aspirated sounds.

- Many speakers have spoken unvoiced aspirated `tha' as voiced aspirated `dha' when `tha' is not in consonant cluster.

- Need for standard test data for Indian languages so that we compare and contrast different systems which is not available now.

- Dearth of speech corpora to work on speech recognition in telephonic and mobile environments.

- If one has to develop real time dictation system, the is need to tackle punctuation, numbers(general and phone numbers), dates etc. which is nontrivial task.

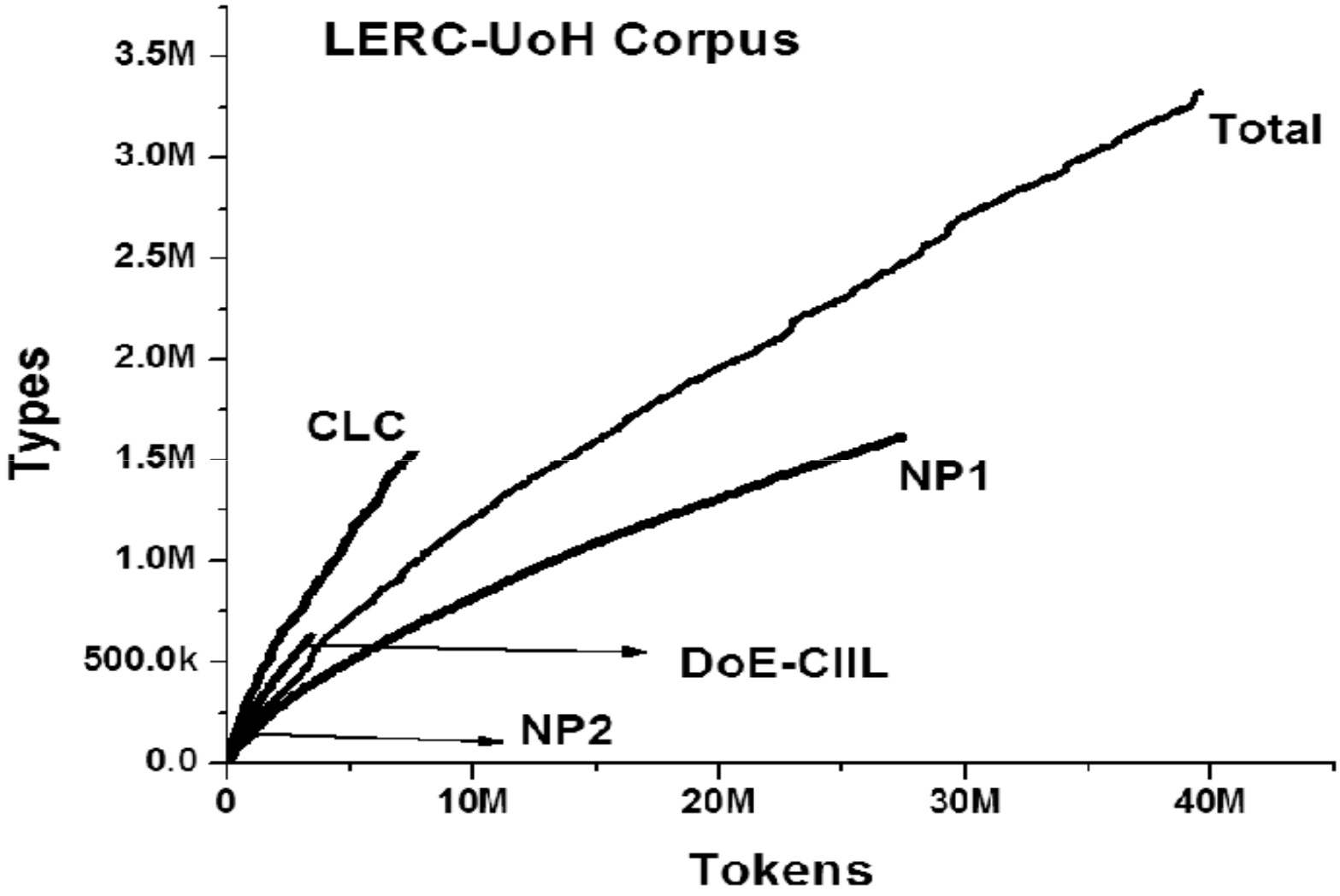
Sub-word models

- G. Lakshmi Sarada et al. group delay based segmentation algorithm for syllable-like units
- Speech recognition performance is about 48.7% and 45.36% for Tamil and Telugu respectively.

- Need good amount of research in this area

Morphological Richness of Telugu

- Telugu language is morphologically very rich because of agglutination.
- External saMdhi (that is, conflation between two or more complete word forms) and compounding add to the numbers.


LERC-UoH Corpus
Types
3.5M
3.0M
2.5M
2.0M
1.5M
1.0M
500.0k
0.0
0
10M
20M
30M
40M
Tokens
Total
CLC
NP1
DoE-CIIL
NP2

Morph Based Language models

- Morph-based language models have been tried for languages such as Finnish, Turkish, and German etc. in recent years and improvement in WER has been observed.

- This aspect has not been studied well in case of Indian languages.

- S. Saraswathi et al. [15] analyzed that the enhanced morpheme-based trigram model with Katz back-off smoothing effect improved the performance of the Tamil speech recognition system

- Word error rate for trigram based models obtained in news and politics domain are 13.8% and 25.04% whereas morph based trigram models gave 12.9% and 23.9% respectively.