Journal of Theoretical and Applied Information Technology
30th June 2012. Vol. 40 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A REVIEW OF SPEECH RECOGNITION WITH SPHINX ENGINE IN LANGUAGE DETECTION

[1]MORCHED DERBALI, [2]MU'TASEM JARRAH, [3]MOHD TAIB WAHID,

[1]Information Systems Department, [2, 3]Information Technology Department,
Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah Saudi Arabia
[3]Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, Malaysia
Tel :[1]+966543430939,[2]+966509952205

Email: [3]mohdtaib.wahid@yahoo.com, [1]morshed@yahoo.com, [2]mutasem.jarrah@yahoo.com

ABSTRACT

Speech recognition is the process of the computer identifying human speech to generate a string of words or commands. The output of speech recognition systems can be applied in various fields. Besides, there are many artificial intelligent techniques available for Automatic Speech Recognition (ASR) development, and hybrid technology is one of it. The common hybrid technique in speech recognition is the combination of Hidden Markov Models (HMMs) and Artificial Neural Networks (ANNs). In this research, Sphinx approach is applied to integrate the advantage of the sequential modeling structure and its pattern classification. Outcome from this paper will assist in next phase of the research which is focusing on building an Arab language speech recognizer by Sphinx4 engine process approach.

Keywords: *Speech Recognition, Automatic Speech Recognition (ASR), Real-Time Operation, Sphinx.*

1. INTRODUCTION

Speech recognition is an area of speech processing that enables humans to communicate with computer through speaking [4]. For a computer to actually understand spoken words, ASR technology concerned with cognitive science and artificial intelligence areas [1]. Researches in ASR always aim to develop techniques that allow computers to recognize in a real-time operation with efficient use of CPU and memory, plus 100% accuracy for all utterances by any person. In order to achieve the goals, limits in ASR such as vocabulary size, noise, speaker characteristics, and channel conditions should be removed. However at this time accuracy greater than 90% is only attained when the task is constrained in some way. Over the years, there are four basic approaches to attain ASR goals [13]:

- **Template-based** approach, where incoming speech is compared with stored units in an effort to find the best match.
- **Knowledge-based** approach that emulate human expert ability to recognize speech.
- **Stochastic or statistical-based** approach, which exploit the inherent statistical properties of the occurrence and co-occurrence of individual speech sounds.
- **Connectionist** approach that use networks of interconnected nodes, which are trained to recognize speech.

Nowadays most speech recognition applications fall into following categories [13]:

- Dictation machines.
- Live transcription of speech for subtitles.
- General dialogue systems such as telephone enquiry systems.
- Command driven "hand-free" operation.
- Security applications such as speaker verification/recognition.
- Archive retrieval (voice mail or video clips).

Figure 1 shows the communication between human and computer via spoken language based on generation and their understanding.

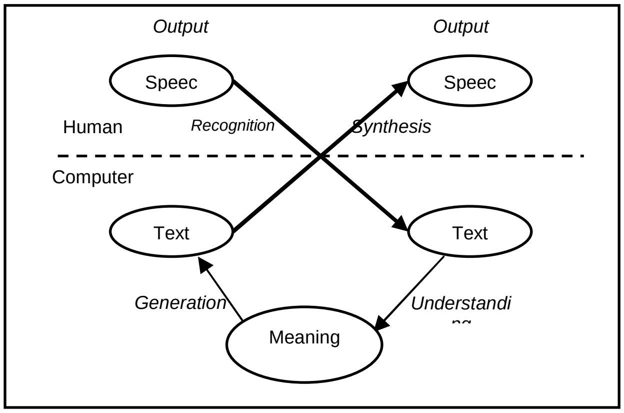

Figure 1: Communication between human and computer via spoken language

Speech recognition systems can be separated in several classes or type depending which modes they are using. Table 1 below shows some of the categorization criteria and its available modes.

Table 1: Type of speech recognition system according to modes used.

Categorization criteria	Available Mode
1. Types of utterances the speech recognition system able to recognize.	▪ Isolated Word Recognition ▪ Continuous Speech Recognition ▪ Spontaneous Speech Recognition
2. Number of speakers able to recognize with great accuracy.	▪ Speaker Dependent System ▪ Speaker Independent System
3. Size of available vocabulary.	▪ Small Vocabulary (10 words) ▪ Large Vocabulary (1000 words)
4. Uses and applications.	▪ Dictation ▪ Command and Control ▪ Telephony ▪ Wearable ▪ Embedded Applications

2. GENERAL HISTORY OF SPEECH RECOGNITION RESEARCH

The research of speech recognition begins four decades ago. In 1952, the first attempt was at Bell Laboratories, where Davis, Biddulph and Balashek built an isolated digit recognition system for single speaker. Most of the early speech recognition systems used spectral information to extract voice features. In the 60's, Japan participated in this area with Suzuki and Nakata, from the Radio Research Laboratories in Tokyo, developed a hardware vowel recognizer in 1961. Meanwhile, in the late 60's, RCA Laboratories worked on the non-uniformity of time scales in speech events. The last development in 60's was the research of Reddy in continuous speech recognition tracking of phonemes at Carnegie Mellon University [12].

In the 70's, researchers achieved number of significant improvements and isolated word recognition became a viable and usable technology. IBM also plays an important role in the area of large vocabulary recognition. Researchers at ATT Bell Labs also conducted a series of experiments, aimed to make speech recognition systems truly speaker independent.

In the 80's, the researchers were shifting from template-based approaches to statistical modeling methods, especially Hidden Markov Models (HMM) [5]. Besides that, neural networks approach was introduced and became very popular in the late 1980's. DARPA efforts to solve large vocabulary and continuous speech recognition problem for defense applications also contribute to this area. In the 90's, a third type of speech recognition architecture has been developed using connectionist techniques.

Nowadays, researchers are focusing on broader area in the speech recognition. The ASR technologies are not only concerned with word content recognition, but also in prosody and personal signature. Addition artificial intelligence qualities such as understanding and learning capabilities are also required in speech recognizer [2][6].

3. FRAMEWORK AND STRUCTURED OF SPEECH RECOGNITION SYSTEM

Speech recognition process or framework in general view is shown below in Figure 2 [6] [12] [13].

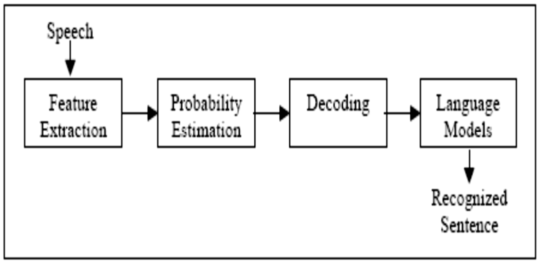

Figure 2: The speech recognition process.

The first stage is record the speaker voice which is consists of acoustic environment and transduction equipment (microphone, preamplifier, filtering, and A/D converter), which effect the generated speech representations. Additive noise, microphone position and type are also part of this component.

The first block, feature extraction is intended to derive acoustic representations that are both good at separating classes of speech sounds and effective at suppressing irrelevant sources of variation.

Then the next two blocks are the core acoustic pattern matching operations in of speech recognition. Probability estimation is the local match, where comparisons are made between speech frames and spectra frames that used for training. As for decoding component, it can be viewed as a global match. The global match is a search for the best sequence of words and is determined by integrating many local matches.

Finally last block consists of language model, which determines the hypotheses that are considered in the global search. This block can further process the global decoder output. If the decoding block generates more than one most likely sentence, language model could re-score the sentence according to grammar or semantics.

Based on research views, speech recognition is a multileveled pattern recognition task, in which acoustic signals are examined and structured into a hierarchy of sub words units (phonemes), words, phrases, and sentences. Structure of a standard of speech recognition system is illustrated in Figure 3 and descriptions of each element are described in Table 2.

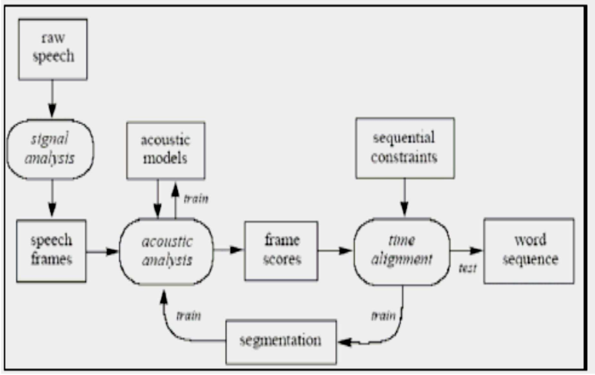

Figure 3: Structure of a standard speech recognition system

Table 2: Elements in standard speech recognition system

Elements	Descriptions
1. Raw speech	• Speech sampled at a high frequency and yields a sequence of amplitude values over time.
2. Signal Analysis	• Transformed and compressed raw speech to simplify subsequent processing. • Popular signal analysis techniques which can extract useful features and compress data without losing important information are Fourier analysis (FFT), Perceptual Linear Prediction (PLP), Linear Predictive Coding (LPC) and Cepstral Analysis.
3. Speech frames	• The result of signal analysis, typically at 10ms intervals, with 16 coefficients per frame. • Provide info of speech dynamics for acoustic analysis.
4. Acoustic models	• For analyzing the acoustic content. • There are many kinds of models, varying in their

Elements	Descriptions
	representation, granularity, and context dependence.
5. Acoustic analysis	• Is performed by applying each acoustic model over each frame of speech to yield frame scores.
6. Frame scores	• For template-based model, score is Euclidean distance between template's frame and unknown frame. • For state-based model, score represents an emission probability. Likelihood current state generated the current frame determined by states parametric/non-parametric function.
7. Time alignment	• The process of searching the best alignment path. • Frame scores are converted to a word sequence by identifying a sequence of acoustic models, which given best total score along an alignment path.
8. Sequential constraints	• Constraints that alignment path must obey, with the fact speech always go forward and never backwards. • Indicating what words may follow what other words.
9. Word sequence	• The end result of time alignment, which is the sentence hypothesis for the utterance.

4. SPEECH RECOGNITION PROCESS AND TECHNIQUE

There are many techniques are used to analyze a speech waveform, among them a few important ones are enumerated below.

A. Oscillogram (Waveform)

Physically the speech signal is a series of pressure changes in the medium between the sound source and the listener. The most common representation of the speech signal is the oscillogram, often called the waveform. In this the time axis is the horizontal axis from left to right and the curve shows how the pressure increases and decreases in the signal [10]. However, a suitable structure is extremely difficult to extract from the mass of information in the intensity waveform. This difficulty motivates us to search for some transformation of the raw intensity waveform into a different representation where the important structure is easier to identify and the enormous amount of variability is reduced. Figure 4 shows the waveform.

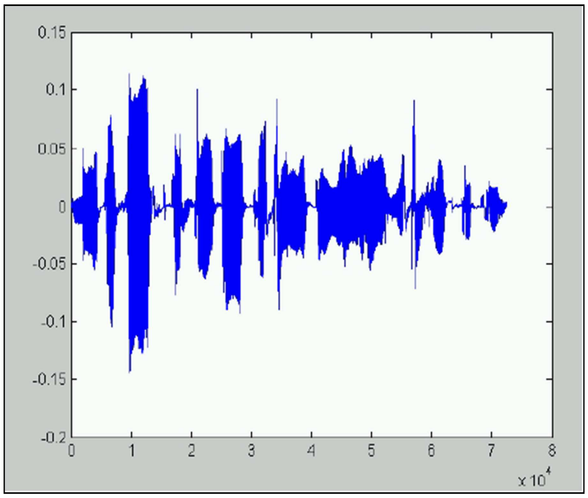

Figure 4: Oscillogram Based View

B. Fundamental Frequency (Pitch)

Another representation of the speech signal is the one produced by a pitch analysis. Speech is normally looked upon as a physical process consisting of two parts: a product of a sound source (the vocal chords) and filtering (by the tongue, lips, teeth etc). The pitch analysis tries to capture the fundamental frequency of the sound source by analyzing the final speech utterance. The fundamental frequency is the dominating frequency of the sound produced by the vocal chords. This analysis is quite difficult to perform. Several algorithms have been developed, but no algorithm has been found which is efficient and correct for all situations. The fundamental frequency is the strongest correlate to how the listener perceives the speaker's accent and stress [10].

C. Spectrum

According to general theories each periodical waveform may be described as the sum of a number of simple sine waves, each with a particular amplitude, frequency and phase. The spectrum gives a picture of the distribution of frequency and amplitude at a moment in time. The horizontal axis represents frequency, and the vertical axis amplitude. If we want to plot the spectrum as a function of time we need a way of representing a three-dimensional diagram, one such representation is the spectrogram. Various speakers have peaks at certain frequencies, resulting in varied speech qualities [10]. Figure 5 shows the Spectrum of a user, speaking the sentence "MICROSOFT, COMPUTER, Triple I T, Yahoo"

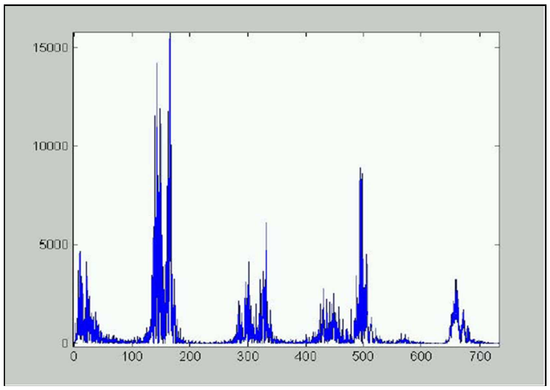

Figure 5: Spectrum of an user, speaking the sentence "MICROSOFT, COMPUTER, Triple I T, Yahoo"

D. Spectrogram

In the spectrogram the time axis is the horizontal axis, and frequency is the vertical axis. The third dimension, amplitude, is represented by shades of darkness. Spectrogram can be considered as a number of spectrums in a row, looked upon "from above", and where the highs in the spectra are represented with dark spots in the spectrogram. From the picture it is obvious how different the speech sounds are from a spectral point of view. The voiced sounds appear more organized. The spectrum highs (dark spots) actually form horizontal bands across the spectrogram. These bands represent frequencies where the shape of the mouth gives resonance to sounds. The bands are called formants. The positions of the formants are different for different sounds [10]. Figure 6 shows a spectrogram for a user speaking the sentence "MICROSOFT, COMPUTER, Triple I T, Yahoo". Note the high amplitudes at the lower end of the frequency ranges. The maximum frequency is 11 KHz for this spectrogram.

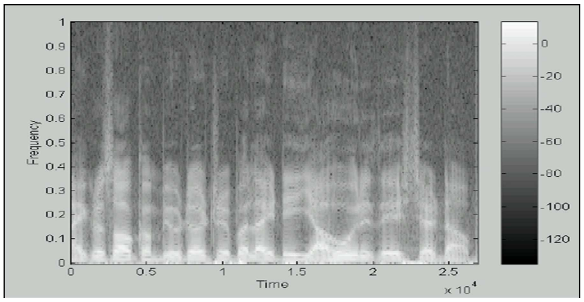

Figure 6: A spectrogram for a user speaking the sentence "MICROSOFT, COMPUTER, Triple I T, Yahoo".

D. Cepstrum

The cepstrum is a method of speech analysis based on a spectral representation of the signal. One way to think of speech is as a signal being filtered by the mouth cavity [10]. Assuming that the actual speech (S) one tries to produce is the same for all people, the signal that comes out of the mouth and into a data recorder is that signal filtered by the person's voice box and throat. If we let v represent this filtering, we can write what we record, r, as the convolution of v and s,

$$r= v*s$$

We need to deconvolve the vocal tract response and the source signal, thus obtaining the fundamental frequency of the speech. If we move the to the frequency domain we would have:

$$R = V S$$

Where R, V, S are the Fourier transforms of r, v, and s respectively. Since we agreed that s is the same for all people, to be able to extract v (or V), there is a need to take the logarithm on both sides to separate the variables.

$$\log R = \log V + \log S$$

Thus, an optimal thing for us to compare from sample to sample is this log R quantity instead of just R because the V and S information are combined additively instead of multiplicatively.

This type of analysis is known as cepstral analysis. As FFT generates both the real and imaginary parts, we only take the magnitude of each FFT component and calculate the logarithm before taking the inverse FFT. For a feature vector x, the real cepstrum 'c' may be calculated by the following formula:

$$[c = \mathrm{real}\ (\mathrm{fft}\ (\log\ (\mathrm{abs}\ (\mathrm{fft}\ (x)))))]$$

5. OVERVIEW AND ARCHITECTURE OF SPHINX

One of the best tools in speech recognition is Sphinx. The Sphinx-4 speech recognition system is a state-of-art HMMs based speech recognition system being developed on open source (cmuSphinx.sourceforge.net) in the Java™ programming language [8]. It is the latest addition to Carnegie Mellon University's repository of Sphinx speech recognition systems [3]. The Sphinx-4 decoder has been designed jointly by researchers from CMU, SUN Microsystems and Mitsubishi Electric Research Laboratories. Over the last few years, the demands placed on conventional recognition systems have increased significantly. Several things are now additionally desired of a system, such as the ability to perform multistream decoding in a theoretically correct manner, with as much user control on the level of combination as possible, that of at least some degree of basic easy control over the system's performance in the presence of varied and unexpected environmental noise levels and types, portability across a growing number of computational platforms, conformance to widely varying resource requirements, easy restructuring of the architecture for distributed processing [3].

The decoder of the Sphinx-4 speech recognition system incorporates several new design strategies which have not been used earlier in conventional decoders of HMM-based large vocabulary speech recognition systems [9]. Some new design aspects include graph construction for multilevel parallel decoding with independent simultaneous feature streams without the use of compound HMMs, the incorporation of a generalized search algorithm that subsumes Sphinx and full-forward decoding as special cases, design of generalized language HMM graphs from grammars and language models of multiple standard formats [8].

The Sphinx-4 architecture has been designed with a high degree of modularity [7][8]. Figure 7 shows the overall architecture of the system. Even within each module shown in Figure 7, the code is extremely modular with easily replaceable functions.

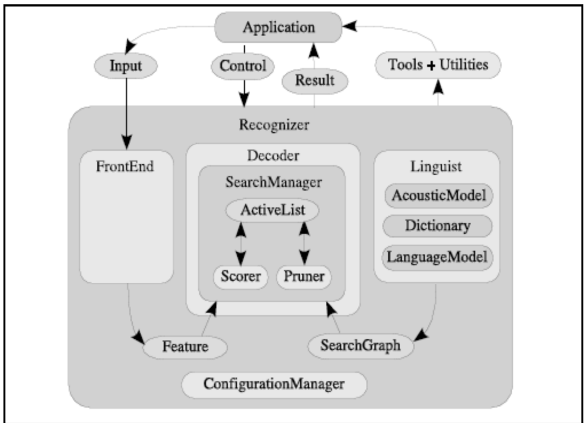

Figure 7 Architecture of the Sphinx-4 system. [7][8]

There are three primary modules in the Sphinx-4 framework: the *FrontEnd*, the *Decoder*, and the *Linguist (Knowledge Base)*. 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 *Acoustic Models*, into a *SearchGraph*. The *SearchManager* in the Decoder uses the Features from the FrontEnd 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. Sphinx-4 also provides support for *Utilities* that support application level processing of recognition results. For example, these utilities include support for obtaining result lattices, confidence scores, and natural language understanding [8].

6. COMPARISON OF DIFFERENT SPEECH RECOGNITION APPROACH

In the previous sections the basic knowledge of HMMs, ANNs and Hybrid HMM/ANN are discussed [9]. In order to know which approach is the best suited to develop a speaker independent continuous speech recognizer, a comparison is made. Table 3 shows the overall comparison among HMM, ANN, and Sphinx.

Table 3: Overall comparisons among HMM, ANN and Sphinx

	ANN	SPHINX	HMM
1. Computational requirement	1	2	2
2. Training complexity	1	3	3
3. Performance	2	3	2
4. Resource demand	2	2	2
5. Vocabulary task	2	2	2

1-Less **2-Average** **3-More**

7. COMPARISON OF DIFFERENT SPEECH RECOGNITION SYSTEM

Speech technology has become the key point of human-machine interface in ELLS. And its research standard is also moving towards practical uses from laboratories. People can get spoken language dialogue between man and machine using speech technology. Many systems, such as booking-ticket automatic question-and-answer system in airport, the tourist automatic question-and-answer system, restaurant reservation automatic consulting system and so on, have achieved good results. Investigation shows that more than 85 percent of people are satisfied with the capability of the information inquiring service system of speech recognition. Inaugurated in 1987 to carry out National 863 Projects, 863 intelligent computer expert groups launched the specific projects for speech technology [11].

Afterwards many national research programs including 985, 973, and 95 projects, national natural science fund and the knowledge innovation projects of Chinese Academy of Science give their support to it. The main manufacturers that develop the Chinese speech recognition include IBM, Microsoft, Speech works, Nuance, Philips, Info talk, Pattek, GR&T, d-Ear Technologies, and English speech recognition include IBM, Dragon Natural Speaking 6 (now Scan Soft), Microsoft, SRI/Nuance Communications = DECIPHER, AT&T Bell Labs (Lucent Tech.), BBN – BYBLOS, CU-HTK, Janus, SPHINX, and Vendors include: Philips, Nuance, Speech Works, IBM, MS, Scan Soft. Table 4 shows the overview of speech recognition system.

Table 4: The comparison main speech recognition systems

System or Organization	**Condition**	**Recognition Performance**	**Description**
CMU SPHINX	Noise background, moderate grammar limitation, continuous speech recognition of 1000 words, speaker-dependent	Recognition rate: 91.1%.	Air Travel Information Service. Its robustness & can deal with various phenomena in spontaneous spoken language effectively and came out in front with the AT&T's CHRONUS & test held by ARPA-ATIS in 1995 since its error ratio is only 3.8%.
	Continuous speech of 997 words under the condition of grammar	Recognition rate: 96.8%, Phoneme recognition rate: 73.8%	
INRS	Speaker-dependent, 75000 words	Recognition rate: 89.5%	-
IBM Tangora (American English)	Speaker-dependent, vocabulary: 5000 words	Recognition rate: 97.1%	It can identify Anglicism, French, German, Italian, Spanish and Japanese
	Speaker-independent, vocabulary: 20000 words	Recognition rate: 94.6%	
IBM Via Voice	Speaker-independent, vocabulary: 32000 Chinese words	Recognition rate: 95%	ViaVoice is the Chinese version of Tangora system
M.Miyatake et al	TDNN synthetically training: 2620 words	Correct rate in searching phoneme: 98.0%	-
	Using predictive HMMs model, 5240 ordinary Japanese words	Recognition rate: 92.6%	
Hild	Speaker-dependent: 1000 sentences, multi-mode TDNN 120 people when speaker-independent: 1680 words	Recognition rate: 98.5%. Recognition rate: 92.0%.	SPHIX: 96.0%. SPHIX: 90.4%.
H.Sawai	A mixed method based on TDNN-LR-DP, 5000 words	Recognition rate: 92.6%	-
K.Iso et al	Predictive HMMs model, speaker-dependent, vocabulary: 5000 words	Recognition rate: 97.6%	It has a strong model- can be used for person- independent continuous speech.

8. CONCLUSION

Connectionist approach has given ASR a new blood. It opened a new road and elicits new investigations in traditional domain of ASR. From the comparison between techniques in speech recognition, Sphinx model is identified as one of the popular connectionist techniques and suitable to use in speech recognition research focusing on building an Arab language speech recognizer. Until this stage of research, it may be concluded that Sphinx is suitable for continuous speech recognition. The big challenge in this research is to develop raw speech database. To execute speech training need more data and variety of continuous speech. Another challenge also came from the language itself, whereas each Arab country has a different dialect and its own set slang expressions. To minimize the problems, for the next phase of this research will focus on the Arabic standard language.

REFERENCES

[1] Matteo Gerosa1, Diego Giuliani1, Shrikanth Narayanan, Alexandros Potamianos, **"**A review of ASR technologies for children's speech", Proceedings of the 2nd Workshop on Child, Computer and Interaction, Cambridge, Massachusetts, USA, Article No. 7, 2009

[2] Chee Peng Lim, Siew Chan Woo, Aun Sim Loh, Rohaizan Osman, "Speech Recognition Using Artificial Neural Networks", 1st Int. Conf. on Web Information Systems Engineering (WISE'00) -Volume 1, Hong Kong, China, June 2000

[3] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, no. 1, pp. 35–45, Jan. 1990.

[4] B. H. Juang and L. R. Rabiner, "Automatic Speech Recognition--A Brief History of the Technology", Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005

[5] X. Huang, A. Acero, and H.-W. Hon. "Spoken Language Processing". Prentice Hall, New Jersey, 2001.

[6] David B. Roe and Jay G. Wilpon, "Voice Communication Between Humans and Machines", National Academy of Sciences (NAS) USA, 1994

[7] Sphinx-4 A speech recognizer written entirely in the JavaTM programming language, http://cmusphinx.sourceforge.net/sphinx4

[8] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition", SML Technical Report Series - 2004-139, Sun Microsystems Laboratories U.S, November 2004.

[9] Steve Young, "A review of large-vocabulary continuous-speech "Signal Processing Magazine, IEEE, Volume 13, Issue 5, Sep 1996 Page(s):45

[10] V. Mantha, R. Duncan, Y. Wu, J. Zhao, A. Ganapathiraju, J. Picone, "IMPLEMENTATION AND ANALYSIS OF SPEECH RECOGNITION FRONT-ENDS", Southeastcon '99 Proceedings IEEE, Lexington, KY, USA, page(s): 32-35,1999.

[11] Jonathan Fiscus William and William M. Fisher and Alvin F. Martin and Mark A. Przybocki and David S. Pallett, "Nist Evaluation Of Conversational Speech Recognition Over The Telephone: English And Mandarin Performance Results", National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, 2000

[12] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon: Spoken Language Processing. Prentice Hall PTR, NJ, 2001

[13] Hesham Tolba & Douglas O'Shaughnessy, "Speech Recognition by Intelligent Machines", IEEE Canadian Review – Summer, 2001