ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Speech Recognition Technology

Naroju Swathi
B.Tech Student,
Sphoorthy Engineering College, Hyderabad.

Mrs.T.Chandrakala
Assistant Professor
Sphoorthy Engineering College, Hyderabad.

Mrs. J. Deepthi, MS, B.Tech
HOD
Sphoorthy Engineering College, Hyderabad.

ABSTRACT

This paper describes recent progress and the author's perspectives of speech recognition technology.

Applications of speech recognition technology can be classified into two main areas, dictation and human-computer dialogue systems. In the dictation domain, the automatic broadcast news transcription is now actively investigated, especially under the DARPA project. The broadcast news dictation technology has recently been integrated with information extraction and retrieval technology and many application systems, such as automatic voice document indexing and retrieval systems, are under development. In the human-computer interaction domain, a variety of experimental systems for information retrieval through spoken dialogue are being investigated. In spite of the remarkable recent progress, we are still behind our ultimate goal of understanding free conversational speech uttered by any speaker under any environment.

This paper also describes the most important research issues that we should attack in order to advance to our ultimate goal of fluent speech recognition.

INTRODUCTION

The field of automatic speech recognition has witnessed a number of significant advances in the past 5 - 10 years, spurred on by advances in signal processing, algorithms, computational architectures, and hardware. These advances include the widespread adoption of a statistical pattern recognition paradigm, a data-driven approach which makes use of a rich set of speech utterances from a large population of speakers, the use of stochastic acoustic and language modeling, and the use of dynamic programming based search methods.

A mechanism of state-of-the-art speech recognizers. Common features of these systems are the use of cepstral parameters and their regression coefficients as speech features, tri hone HMMs as acoustic models, vocabularies of several thousand or several ten thousand entries, and stochastic language models such as bigrams and trigrams. Such methods have been applied not only to English but also to French, German, Italian, Spanish, Chinese and Japanese.

Although there are several language specific characteristics, similar recognition results have been obtained.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Speech Processing

Signal processing:

- Convert the audio wave into a sequence of feature vectors

Speech recognition:

- Decode the sequence of feature vectors into a sequence of words

Semantic interpretation:

- Determine the meaning of the recognized words

Dialog Management:

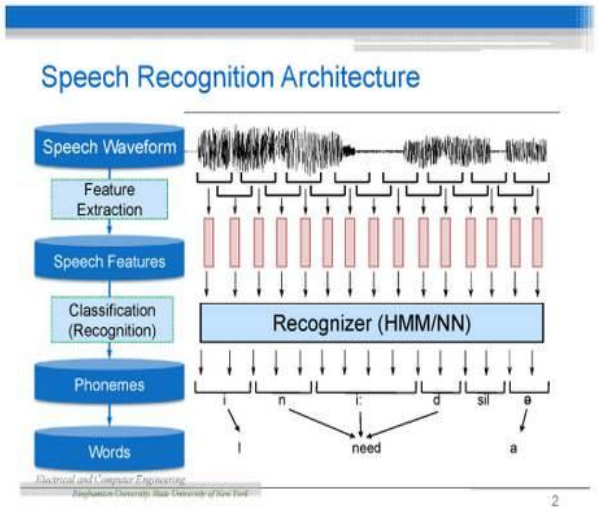
- Correct errors and help get the task done

Response Generation

- What words to use to maximize user understanding

Speech synthesis (Text to Speech):

- Generate synthetic speech from a 'marked-up' word string



The speech recognizer uses n-gram back off language models estimated on the transcriptions of spoken queries. Since the amount of language model training data is small, some grammatical classes, such as cities, days and months, are used to provide more robust estimates of the n gram probabilities. A confidence score is associated with each hypothesized word, and if the score is below an empirically determined threshold, the hypothesized word is marked as uncertain.

A Multimodal Dialogue System for Information Retrieval

We have recently investigated a paradigm for designing multimodal dialogue systems. An example task of the system was to retrieve particular information about different shops in the Tokyo Metropolitan area, such as their names, addresses and phone numbers. The system accepted speech and screen touching as input, and presented retrieved information on a screen display or by synthesized speech as shown in Fig. 6. The speech recognition part was modeled by the FSN (finite state network) consisting of keywords and fillers, both of which were implemented by the DAWG (directed acyclic word-graph) structure. The number of keywords was 306, consisting of district names and business names. The fillers accepted roughly 100,000 non-keywords/phrases occurring in spontaneous speech. A variety of dialogue strategies were designed and evaluated based on an objective cost function having a set of actions and states as parameters. Expected dialogue cost in

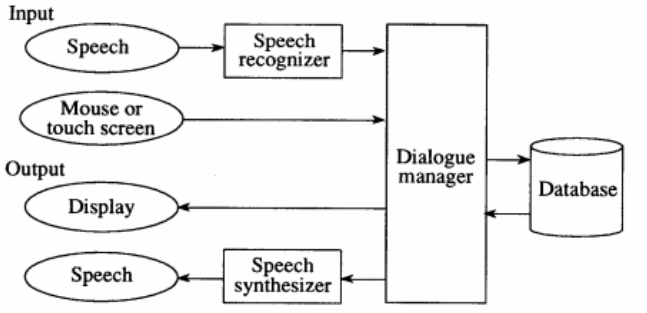


Fig. 6 - Multimodal dialogue system structure for information retrieval.

Language modeling for spontaneous speech recognition

One of the most important issues for speech recognition is how to create language models (rules) for spontaneous speech. When recognizing spontaneous speech in dialogues, it is necessary to deal with variations that are not encountered when recognizing speech that is read from texts.

These variations include extraneous words, out-of-vocabulary words, ungrammatical sentences,

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

disfluency, partial words, repairs, hesitations, and repetitions. It is crucial to develop robust and flexible parsing algorithms that match the characteristics of spontaneous speech. A paradigm shift from the present transcription-based approach to a detection-based approach will be important to solve such problems. How to extract contextual information, predict users' responses, and focus on key words are very important issues.

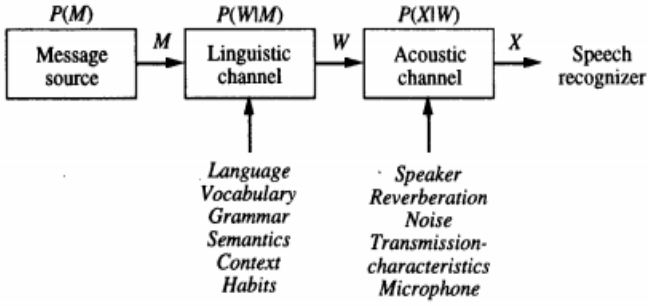
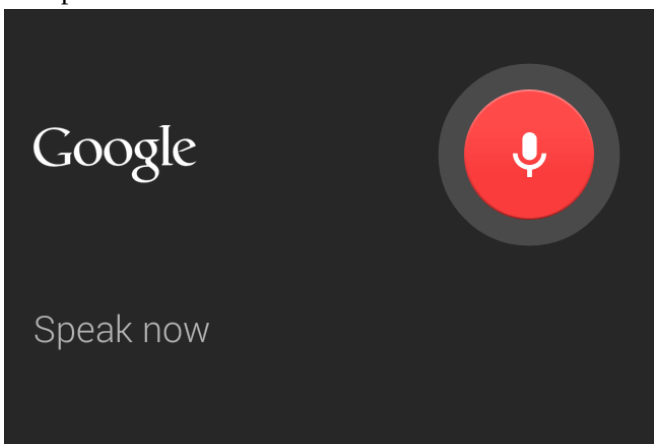


Fig. 9 - A communication - theoretic view of speech generation and recognition.

CONCLUSIONS

Speech recognition technology has made a remarkable progress in the past 5 - 10 years. Based on the progress, various application systems have been developed using dictation and spoken dialogue technology. One of the most important applications is information extraction and retrieval. Using the speech recognition technology, broadcast news can be automatically indexed, producing a wide range of capabilities for browsing news archives interactively.

Since speech is the most natural and efficient communication method between humans, automatic speech recognition will continue to find applications, such as meeting/conference summarization, automatic closed captioning, and interpreting telephony. It is expected that speech recognizer will become the main input device of the "wearable" computers that are now actively investigated. In order to materialize these applications, we have to solve many problems. The most important issue is how to make the speech recognition systems robust against acoustic and lingustic variation in speech. In this context, a paradigm shitt from speech recognition to understanding where underlying messages of the speaker, that is, meaning/context that the speaker intended to convey are extracted, instead of transcribing all the spoken words, will be indispensable.



REFERENCES

[1] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs", Proc. Int. Conf. Spoken Language Processing, Banff, pp. 369-372 (1992)

[2] S. Furui: "Recent advances in robust speech recognition", Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 11-20 (1997)

[3] J. Choi, D. Hindle, J. Hirschberg, I. MagrinChagnolleau, C. Nakatani, F. Pereira, A. Singhal and S. Whittaker: "SCAN - speech content based audio navigator: a systems overview", Proc. Int. Conf. Spoken Language Processing, Sydney, pp. 2867-2870 (1998)