CSME 206A Natural Language & Speech Processing **Spring Semester**

Lecture 13: Introduction to Automatic Speech Recognition

Lecturer: K .R. Chowdhary *: Professor of CS*

13.1 Introduction

Speech has long been viewed as the future of computer interfaces, promising significant improvements in ease of use over the traditional keyboard and mouse. In the past few decades, dramatic improvements in speech recognition technology have made high-performance algorithms and systems available. These advancements have been supported by government-sponsored scientific research in industrial research labs and in Universities.

Despite this long history of research progress, speech technology was, at least initially, slow to find commercial application. However, in the past couple of years the market has begun developing rapidly, and several interesting and useful speech applications are now feasible.

We can consider applying speech recognition in either of two primary modes: using speech as 1. spoken input or, 2. as data / knowledge source. Spoken input addresses applications like dictation systems and navigation or transactional systems. With dictation applications, the system transcribes spoken words verbatim into written text. These applications can create text such as personal letters, business correspondence, or even e-mail messages. In transactional applications, users can navigate around the application or use speech to conduct a transaction. For example, speech can be used to purchase stock, reserve an airline itinerary, or transfer bank account balances.

In other type of application, i.e., using speech as a data/knowledge source paves the way for applications such as meeting capture and knowledge management. These applications begin modestly as multimedia indexing systems that use speech recognition to transcribe words verbatim from an audio file into text. Subsequently, information retrieval techniques applied to the transcript create an index with time offsets into the audio. Users can access the index using text keywords to search a collection of audio or video documents [1].

The task of speech recognition or Automated Speech Recognition (ASR) is a process to convert speech into a sequence of words through a computer program. As the most natural communication modality for humans, the ultimate aim of speech recognition is to enable human to communicate more naturally and effectively with computers. While, the long-term objective of speech recognition requires its integration with many NLP (Natural Language Processing) components.

There are many emerging applications that can be readily deployed with the core speech-recognition module. Some of these applications include voice dialing, call routing, data entry and dictation, command and control, and computer-aided language learning. Most of these modern systems are based on *statistic models* such as *Bayes Probability* and *Hidden Markov models* (HMMs). One reason why HMMs are popular is that their parameters can be estimated automatically from a large amount of data, and they are simple and computationally feasible [3].

Speech recognition is often regarded as the front-end for many NLP components. In practice, the speech

system typically uses context-free grammar (CFG) or statistic n-grams[1] for the same reason that HMMs are used for acoustic modelling.

The general problem areas in speech recognition are *graph searching* and *automata manipulation*. Some tight theoretical bounds and practical implementations for some problems, like, shortest path finding and finite state automata minimization, are well known manifestations of these problems. The major part of the progress in speech recognition is due to good heuristic methods that solve special cases of the general problems. However, practical implementations of any algorithm is critical to the deployment of speech recognition technology [2].

13.2 Scope and limitations of Speech Recognition

The Spoken language is effective for human-human interaction but often has severe limitations when applied to human-computer interaction. Speech is slow for presenting information, it is transient, therefore difficult to review or edit later. It interferes significantly with other cognitive tasks of human, like, actions, vision, and thinking. For example, when one is attentively listening to someone, he/she cannot plan a shopping schedule [4].

However, the speech has many interesting features. The speech has been proved to be useful for storing-and-forwarding messages, it alerts in busy environments, and has been found to be useful for input-output for blind, and for motor-impaired persons. Speech recognition and generation is sometimes helpful for environments where our hands are busy, eyes-busy, or our mobility is necessary.

Obvious physical limitation of human-based speech includes fatigue from speaking continuously, and disruption in an crowd where many people are speaking.

If we analyse the differences between human-human interaction vs human-computer interaction, based on the inferences drawn it may be possible to choose appropriate applications suitable for human-computers interactions. The key distinction amongst these two is: rich emotional content in human speech conveyed in the form of *prosody*[2], *pacing*[3], *intonation*[4], and *amplitude*.

13.2.1 Multitasking by Brain

Now consider human *acoustic memory* – a short-term and working memory (also called verbal memory). Part of the human brain transiently holds chunks of information and solves problems, also supports speaking and listening. Therefore, working on tough problems is best done in quiet environments–without speaking or listening to someone, because, speaking and listening cannot be done synchronously with thinking or reasoning.

However, because physical activity is handled in another part of the brain, problem solving is compatible with routine physical activities, like walking and driving. In short, humans can easily speak and walk both together, but find it difficult to speak and think.

[1] n-gram: In computational linguistics, an n-gram is continuous sequence of n-items from a given sample of text or speech. The items can be phonemes, syllables, letters, words, or base pairs according to application. The n-grams are typically collected from a text or speech corpus.

[2] prosody: the rhythmic structure of versed text,

[3] pacing: the speed at which a story is told,

[4] intonation: rise and fall of the voice pitch,

Similarly, when operating a computer, most humans type using keyboard (or move a mouse) and think but find it difficult to speak and think together. The hand-eye coordination is accomplished in different structures of brain, so typing or mouse movement can be performed in parallel with problem solving (thinking).

In summary, the number of speech interaction success stories are increasing slowly; the designers should conduct empirical studies to understand the reasons for their success, as well as their limitations, and alternatives. A particular concern for everyone today, while driving, is a plan by several automobile manufacturers is to introduce email handling via speech recognition for automobile drivers, there is already convincing evidence of higher accident rates by smart-phone users. This is due to particularly, those drive and handle calls.

Realistic goals for speech-based human-computer interaction, can be better understood through knowledge of human multitasking models, and by understanding of how human-computer interaction is different from human-human interaction.

13.2.2 Near future Applications

On the positive side, it is expected that speech messaging (already common), alerts, and input-output are to grow in popularity in their usage. Dictation designers will find useful niches, especially for routine tasks. There will be happy speech-recognition users, such as those who wish to quickly record some ideas for later review. Telephone-based speech recognition applications, such as voice dialing, directory search, banking, and airline reservations, will be useful complements to graphical user interfaces. But, for many tasks, we see more rapid growth of reliable high-speed visual interaction over the Web, as a likely scenario. Similarly for many physical devices, carefully engineered control sticks and switches will be effective while preserving speech for human-human interaction and keeping rooms pleasantly quiet.

13.3 Challenges due to Variability of speech sounds

Unlike the printed text, there are no well defined boundaries between phonemes (and even words) due to many important facts of continuous speech. For instance the phrase "six sheep" may easily confused with "sick sheep". There are external parameters that can effect the performance of the speech recognition system, these are characteristics of environment noise, type of the microphone, and placement of the microphone.

The acoustic realization of phonemes, the smallest unit by which the words are composed, are highly dependent on the context in which they appear. These *phonetic variabilities* are exemplified by the acoustic differences of the phonemes, the /t/ in *two*, *true*, and *butter*, are the examples, where t is pronounced differently. At word boundaries the contextual variations can be very high, for example *gas shortage* may sound like *gash shortage*.

The acoustic variabilities can occur due to change in environment, in position, and speaker's physical and emotional taste, speaking rate, or voice quality. Finally, differences in speaker's socio-linguistic background, dialect, and vocal tract size and shape can contribute to across speaker variabilities.

Following are the factors due to unclear boundaries in speech [5].

- *Physiological.* Speech waveforms of a vowel may very due to different pitch frequencies, as well due to different dimensions of vocal tract. The latter leads to different resonance frequency of oral cavity. The resonance frequencies of male adults are in general lower than female in same age group.

- *Behavioral.* The speaking rate (words per unit time) vary among the persons. The accent and usage of words depend on region and social back-ground of speaker. Pronunciation of unfamiliar words often deviates the pronunciation from standard.

- *Transducer.* The quality of the transducer, its frequency response (variation of gain with frequency) also change the sound to be recognized. Also, the distortions in microphone, effects the speech to be recognized.

- *Environment conditions.* Presence of background noise, background speech of neighbours, also effect the recognition process.

- *Phonetic context.* The acoustic manifestation of speech sound of a word or syllable, depends a lot on the preceding and following sounds. This is due to inertia of articulators (called co-articulation). The context dependent variability of sound is integral part of the systems, and governed by certain rules, hence these sounds can be modelled by detailed phonetic units for recognition. For example, in the phrase "flight to New Delhi", the letter sound /t/ of "flight" and /t/ in "to", are continuous, hence they may appear continuous due to co-articulation if the pronunciation is carried out at speed (see Fig. 13.1).

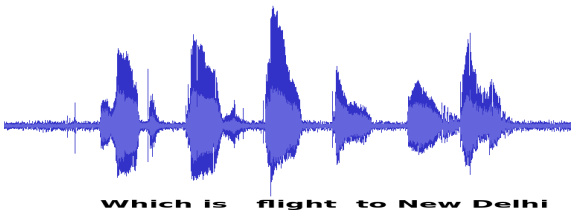

Figure 13.1: Isolated word recognition in sentence: "Which is flight to New Delhi."

13.4 Phases of Speech recognition

The aim of speech recognition is to reconstruct the original text of a spoken sentence from continuous acoustic signal induced by the associated utterances. A speech recognizer usually operates in phases, as sown in Fig. 13.2, referred to as *recognition cascade*. Using signal processing, the acoustic waveform is first transformed into a sequence of discrete observations over some unbounded alphabet[5] $\mathcal{F}$. We call the sequence of discrete observations as *observation* or *input sequence*. Its symbols are referred to as *feature vectors*, which are designed to preserve the relevant information from the original signal. Obtaining the feature vectors overcome the problem of variability of speech due to factors discussed above. In addition, the effect of reducing the speech signal feature vector is to reduce the storage size, as the feature vectors consume far less space than the original signal, thus providing the compression of information.

In general, the feature vectors also have probability distribution associated with them. The probability distribution can also be taken as continuous instead of discrete. The issue of signal processing is equally important here. However, in this text we will be discussing the acoustics issues which are necessary for understanding of computational aspects, as well the computational aspects of speech processing and not cover the signal processing aspects [2].

[5]No precise boundaries when alphabets are pronounced as part of a word.

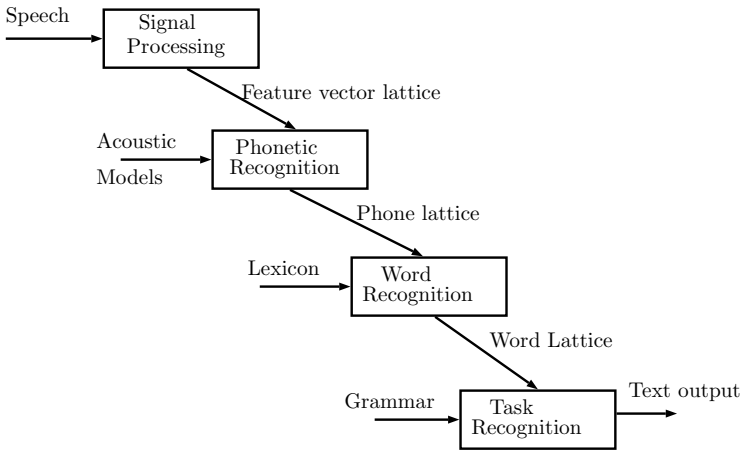

Figure 13.2: Phases in speech recognition.

The phonetic recognition phase produces phonemes at the output, while its inputs are feature vectors produced by first phase and the acoustic models of phonemes. To recognize the words, the lexicon models (i.e., dictionary) are input along with the phone lattice produced by previous phase.

The final phase has input the word lattice, and grammar, using which it recognizes the sentences and provides final text as output.

Since different persons can utter the same sentence in different ways, and similar is the case with same person uttering the same sentence differently. This motivates us to consider the process of recognition as stochastic.

At the more advanced level of speech recognition, we can divide it in two parts: 1. recognition of isolated words (for example "speech" and "recognition" in the phrase "speech recognition"), and 2. recognition of continuous speech – sequence of words spoken at speed.

13.5 Search Algorithms

There are basically two types of search algorithms: 1. Isolated word recognition (IWR) and 2. Continuous word recognition (CWR).

13.5.1 Isolated word recognition

In the IWR algorithm, where the recognizer takes as input one word at a time, isolated with a small time gap, the words belonging to some dictionary, and output with high probability the word which has been actually spoken. The algorithm for isolated words is nothing but *lexicon* algorithm.

The IWR search algorithm compares the input sequence to the canonical acoustic model for each word in the lexicon, and output the word that maximizes a given objective function. The objective function is the *likelihood* of a word, given the observation sequence. In practice, the computation of the objective function is usually approximated using heuristics, whose effectiveness is established experimentally. Such an approximation is justified due to availability of fast response and availability of large dictionaries and lexicons.

13.5.2 Continuous word recognition

For CWR, the *search* is implemented as in finite automata. The lexicon database contains typical pronunciations of each word in the dictionary, e.g., for English language it is set of phonetic transcriptions of words of English Dictionary. From phonetic transcriptions, one can obtain *canonical acoustic* models for the words in the dictionary. These acoustic models can be considered to be Markov sources over the alphabet $\mathcal{F}$.

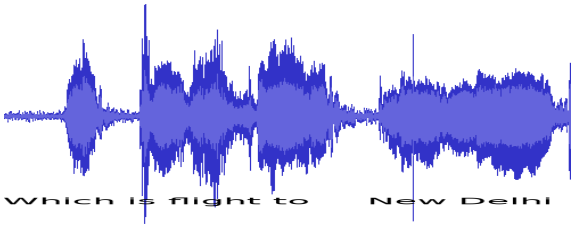

Figure 13.3: Continuous words in sentence: "Which-is-flight-to-New-Delhi."

In CWR, the recognizer takes as input the observation sequence corresponding to a spoken sentence, and outputs with high probability the sentence, that might have caused the utterance.

The three algorithmic components of of CWR system :

1. *lexicon*,
2. *language model*, and
3. *search algorithm*.

The lexicon is basically what we have in isolated word recognition, while the language model gives a stochastic description of the language. This description is a syntactic description, in addition to probabilistic description. As per the syntactic description, the specific words will follow specific word(s) in any given sentence. For example, which specific noun will follow the specific verb "enters", in a word sequence. For example, "The teacher enters in class", is more likely than "The teacher enter the glass !".

The lexicon is obtained as in the case of isolated word recognition, whereas the language model is built using linguistic as well as task-specific knowledge. The search algorithm uses the language model, where each word in acoustic models is derived from the lexicon to "match" the input sequence. The matching tries to find a grammatically correct sentence that maximizes a given objective function. Like in objective function, one would like to have the maximum *likelihood* of a sentence, given the observation sequence.

13.6 Challenges of Continuous Word Recognition

In the begin, the isolated word recognition scheme appears as a special case of continuous word (speech) recognition, and hence, algorithm design may turn out to be easy. But, this is not the case, as the nature of search procedure differs heavily in the two cases. Obviously, identification of letters in a words can be done in similar way as identification of words in a sentence. Here is more technical reason: In isolated word recognition case, all the needed acoustic information for modeling the words in the dictionary is available to the search procedure. That is, for each word in the dictionary, there is a canonical acoustic model of

that word in the lexicon. Thus, the search problem is nothing but pattern recognition, because, the search procedure tries to find the canonical model that best matches the input observations.

However, for continuous speech recognition, the acoustic information that models a sentence in the language is given only partially and implicitly in terms of rules. There does not exists such canonical acoustic model for each sentence in the language. The only canonical acoustic models that are available, are those in lexicon that corresponds to the words in the language. The search procedure must therefore assemble a sequence of canonical acoustic models that best match the observation sequence, guided by the rules of the language. However, such an assembly become complicated due to inter-word dependencies, i.e., when we utter a sentence, the sounds associated with one word influence the sounds associated with next work due to *co-articulation* effect of each *phone* on successive phones[6]. These inter-word dependencies are not completely modeled and described by the lexicon and other language models, because, otherwise there is need of a separate canonical model for each sentence. Resulting to these facts, the search procedure for continuous speech recognition faces additional challenge – using incomplete information – where actually a word begins and ends? In fact, for a given observation sequence, the search procedure usually postulates many different word boundaries, which may in turn lead to exponentially many ways of decoding the input sequence into a sequence of words [2] !

Fig. 13.1, and 13.3 demonstrate the difference between IWR and CWR. Each figure displays acoustic waveform and labeling of the sentence, "Which is flight to New Delhi." In Fig. 13.1 words are spoken in isolation, but in Fig. 13.3 the sentence is spoken fluently. The acoustic waveforms displays the signal amplitude as a function of time. The acoustic waveforms is hand-segmented into phones as well into words. We note that in isolated words it contains distinct and easy-to-detect boundaries without co-articulation effect on phone boundaries. In continuous speech, however, the word boundaries are not clear, and the phones at word-boundaries display co-articulation effect. In the extreme case, the /t/ of "flight" and /t/ of "to" have joined. Hence, in CWR, not only we face the problem of finding the word boundaries, but also the due to co-articulation effect, the acoustic models for each word in lexicon does not fit necessarily reflect the actual utterance of the word in the sentence. Though the Fig. 13.3 appears compressed in time scale, however, both the figures are in same time scale, in 1st, the words are distinctly (or slowly) pronounced, while in Fig. 13.3 the sentence is spoken with speed.

13.7 Parameters of ASR

A speech recognition system can be characterized by many parameters, some of the more important parameters are given in table 13.1.

An isolated word recognition system requires the speaker to briefly pause between the words, where as the continuous word recognition does not. Some system require speaker enrolment, where some some samples of speaker are needed to be provided to system before using the system. Where as other speech recognition systems are speaker independent, where no speaker enrolment is necessary.

When speech is provided by sequence of words, language models or artificial grammars are used to restrict the combination of words. A simplest language model can be specification of finite-state machine network, in which permissible words following to each word are explicitly given. The general language models approximating natural language are specified in terms of context-sensitive grammars [6].

One popular measure of the difficulty of task is – combining the vocabulary size and the language model,

[6]Consider the utterances: "How to recognize speech," vs "How to wreck a nice beach," where, due to overlapping of articulation, two different sentences give the impression, as if identical sentences have been uttered.

Table 13.1: Typical parameters used to characterize the capability of ASR.

Parameter	Range
Speaking mode	Isolated words to continuous speech.
Speaking style	Read the speech to continuous speech.
Environment	Speaker-dependent to speaker-independent.
Vocabulary	Small (less than 20 words) to large (more than 20,000 words).
Language model	Finite-state to context-sensitive.
Perplexity	Small (less than 10) to large (more than 100)
Signal-to-noise ratio	High (more than 30 dB) to low (less than 10 dB).
Transducer	Voice-cancelling microphone to telephone.

called *perplexity*, which approximately means geometric mean of the number of words that can follow a word after language model has been applied.

Review Questions

1. What can be the major applications of Speech recognition?
2. What can be the major applications of TTS?
3. Which of the following operation pairs are supported parallelly by the brain?
   (a) Speaking and reasoning
   (b) Walking and talking
   (c) Talking and thinking
   (d) Thinking and listening
   (e) Walking and listening
   (f) Typing and thinking
   (g) Mouse movement and thinking
   (h) Mouse movement and talking
4. Why the speech takes more space than the text of same speech?
5. What are the immediate near future applications of automatic speech recognition?
6. What are the external parameters, that can be hurdle for speech recognition?

Exercises

1. Out of the major applications of speech recognition, which are the applications due to World-Wide-Web?

2. What are technical reasons due to which the speech sound varies:
   (a) From person-to- person for the same sequence of words?
   (b) Same word repeated by the same person?
   (c) Due to age factor for same person?
   (d) Due to male / female voice?
3. What are the features in speech due to which we can recognize the speaker?
4. Describe Isolated Word Recognition (IWR) algorithm in your own words.
5. Describe Continuous Word Recognition (CWR) algorithm in your own words.
6. Explain the phases of speech recognition in detail, with the help of block diagram.
7. Explain the significance of "language model", and "lexicon model", for automatic speech recognition.
8. What functions (speech, reasoning, physical activities) can be performed in parallel by the multitasking brain? Give your answer based on some inferences.
9. The speech for, say, "Hello world", has more information contents than the text "Hello world". Explain, why?
10. What are the telephone-based speech recognition applications?
11. What do you understand by "acoustic variabilities"? What are the causes of their occurrence?
12. How the understanding of speech recognition by human can be useful for implementation of speech recognition by machine?
13. In your view, what may be the resources needed for speech recognition? Do you know, what is a "speech engine"? What are its functions? Is it any way related to "search engine"?

References

[1] D. Jurafsky and J. Martin, *Speech and Language Processing, 3rd Ed.*, Pearson India, isbn: 3257227892, Nov. 2005.

[2] Buchsbaum, Adam L. and Giancarlo, Raffaele, *Algorithmic Aspects in Speech Recognition: An Introduction*, J. Exp. Algorithmics, jan. 1997, issn:1084-6654, `http://doi.acm.org/10.1145/264216.264219`.

[3] Indurkhya, Nitin and Damerau, Fred J., *Handbook of Natural Language Processing, 2nd ed.*, isbn: 1420085921, 9781420085921, Chapman & Hall/CRC, 2010.

[4] Shneiderman, Ben, *The Limits of Speech Recognition*, Commun. of the ACM, Vol. 43, No. 9, Sept. 2000, pp. 63–65.

[5] Samudravijaya K, *Automatic Speech Recognition*, `http://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/`, urldate = 2018-04-30.

[6] Cole and Ron, *Survey of the State of the Art in Human Language Technology*, isbn: 0-521-59277-1, Cambridge University Press, New York, NY, 1997.