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

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 [ID10].

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 for the same reason that HMMs are used for acoustic modeling.

The general problem areas in speech recognition are *graph searching* and *automata manipulation*. In addition, 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 [BA97].

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, she or he cannot plan a shopping schedule [Shn00].

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 analyze 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*[1], *pacing*[2], *intonation*[3], 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 that 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. 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.

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.

[1] prosody: the rhythmic structure of versed text,
[2] pacing: the speed at which a story is told,
[3] intonation: rise and fall of the voice pitch,

13.3 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". Following are the factors due to unclear boundaries in speech [**?**]:

- **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 neighbors, 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 modeled by detailed phonetic units for recognition. For example, in the phrase "flight to New Delhi", the letter sound /t/ of "flight" and of /t/ in "to", are continuous, hence they may appear continuous due to co-articulation if the pronunciation is carried out at speed.

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 figure 13.1, referred to as *recognition cascade*. Using signal processing, the acoustic waveform is first transformed into a sequence of discrete observations over some unbounded alphabet $\mathcal{F}$. We call the sequence of discrete observations the *observation* or *input sequence*. Its symbols are referred to as *feature vectors*, which are designed to preserve the relevant information from the original signal. 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. 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 [BA97].

Since different persons can utter the same sentence at same time in different ways, and same is the case with same person uttering the same sentence differently. This motivates to consider the process of recognition as stochastic. At the more advanced level we can divide the speech recognition in two parts: 1. recognition of isolated words (for example "speech" and "recognition" in the phrase "speech recognition"), and 2. recognition of continuous speech (for example, recognition of word "speech"), where we recognize the structure of a word.

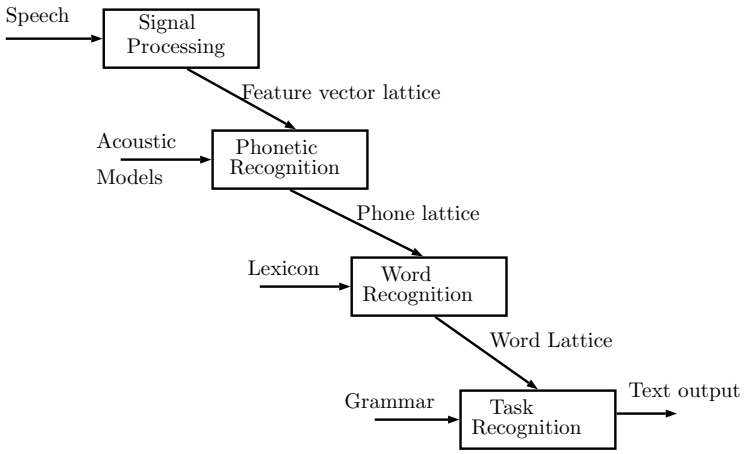

Figure 13.1: Phases in speech recognition.

13.4.1 Search Algorithms

One type of algorithm is called isolated word recognition (IWR), 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. For continuous word recognition the *search* 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 $\mathcal{F}$.

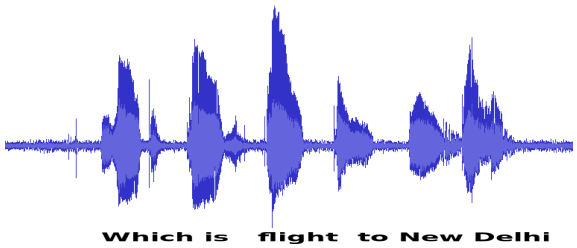

Figure 13.2: Isolated word recognition in sentence: "Which is flight to New Delhi."

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, give the observation sequence. In practice, the computation of the objective function is usually approximated using heuristics, the effectiveness of which is established experimentally. Such an approximation is justified due to the need of fast response in the presence of large dictionaries and lexicons.

In continuous word recognition (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 the recognizer are: *lexicon*, the *language model*, and *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 that, specific words will follow other word(s). For example, which specific noun will follow the specific

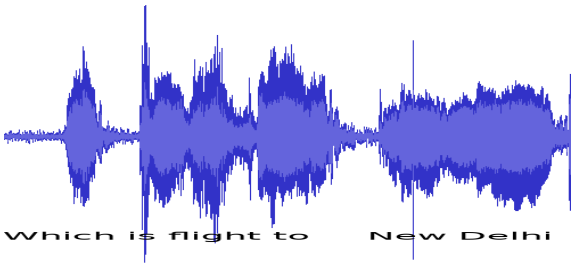

Figure 13.3: Continuous words in sentence: "Which-is-flight-to-New-Delhi."

verb "enters", in the word sequence: "The teacher enters the ....", the answer "class" is more likely than "laboratory".

The lexicon is obtained as in 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 and, for each word, the acoustic models derived from the lexicon, to "match" the input sequence, trying to find a grammatically correct sentence that maximizes a given objective function. As an objective function, ere again one would like to use the maximum *likelihood* of a sentence given the observation sequence.

13.4.2 Co-articulation effect

At first, 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 because 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, as the search procedure tries to find the canonical model that best matches the input observations.

In continuous speech recognition, the acoustic information modeling the sentence in the language is given only partially and implicitly, in terms of rules. There is no 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 is complicated by 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[4]. 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 intern lead to exponentially many ways of decoding the input sequence into a sequence of words [BA97] !

Figure 13.2, and 13.3 demonstrate the difference between IWR and CSR. Each figure displays acoustic waveform and labeling of the sentence, "Which is flight to New Delhi." In figure 13.2 words are spoken in isolation, but in figure 13.3 the sentence is spoken fluently. The acoustic waveforms displays the signal

[4]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.

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 one face the problem of finding the word boundaries, but also the due to co-articulation effect, the acoustic models for each word in lexicon do not necessarily reflect the actual utterance of the word in the sentence. Though the figure 13.3 appears compressed in time scale in time scale, however, both the figures are in same time scale, in 1st, the words are distinctly pronounced, while in Figure 13.3 the sentence is spoken with flow.

Exercises

1. What are the potential applications of speech recognition, which are due to WWW?
2. How the speech recognition can be helpful to disables?
3. What are the immediate near future applications of automatic speech recognition?
4. What are technical reasons due to which the speech sound various from person to person for the same sequence of words?
5. What are the factors due to which the speech of any two persons cannot be identical?
6. What are the factors due to which we are able to recognize the speaker?
7. Write the Isolated Word Recognition algorithm in your own words.
8. Write the Continuous Word Recognition algorithm in your own words.

References

[BA97] Giancarlo R Buchsbaum AL. Algorithmic aspects in speech recognition: An introduction. *Journal of Experimental Algorithms*, 1997.

[ID10] Nitin Indurkhya and Fred J. Damerau. *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2nd edition, 2010.

[Shn00] Ben Shneiderman. The limits of speech recognition. *Commun. ACM*, 43(9):63–65, September 2000.