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

Lecture 7: Markov Chains for Speech Recognition

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

7.1 Models and Speech Recognition

The ASR (automatic speech recognition technology) has matured in recent years only in the human-computer interaction, but the other applications are still far from reality. The most successful application is in telephones, for complaint registering with telephone company. The ASR is also being applied in dictation, for monologue by a specific speaker. The dictation is common in law, for documenting the debate before it is argued in the court of law.

Different applications of speech require different constraints, and hence require different algorithms. We assume that input text words are separated by small pauses. We will introduce the modern speech recognizer components: the Hidden Markov Model (HMM), the idea of spectral features, and some new algorithms.

The speech recognition system treats the acoustic input as noisy channel, i.e, the sentence available is a noisy version of original sentence. In order to "decode" this noisy sentence, we consider all possible sentences and then find out the probability that it is the original sentence. For this, choose the one having maximum probability of likely-hood. This approach is shown in figure 7.1.

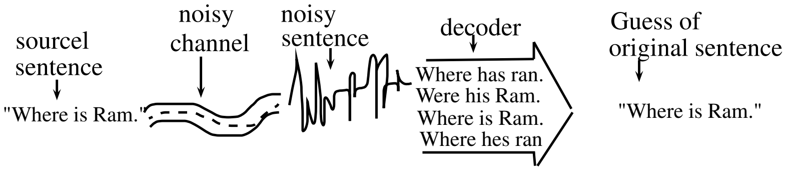

Figure 7.1: Noisy channel model.

The modern speech recognition works by searching huge space of potential "source" sentence, and choosing the one which having the highest probability of generating the noisy sentence. This requires the solution of two problems:

- Complete metric for best match of noisy and original sentence. This is because these two sentences can never match exactly. Hence, we make use of probability.
- English has large number of sentences, hence use some search technique, which does not require to search the entire space, but still can locate the current correct sentence.

We can treat the acoustic input O as a sequence of individual symbols (e.g. slices of inputs $(o_1, o_2, o_3, \dots)$, say one symbol for every 10 milli-secs., and represent each slice by *frequency* or *energy* of that slice. Let us assume,

$$O = o_1, o_2, \ldots o_t. \tag{7.1}$$

Let the corresponding sentence, which might have caused this sentence, is a string of words:

$$W = w_1, w_2, \ldots, w_n \tag{7.2}$$

The probabilistic implementation of the above mentioned intuition is. i.e., probability of occurrence of this sentence, given that acoustic input O is evidenced, is

$$\hat{W} = argmax_{w \in \mathcal{L}}\ P(W|O) \tag{7.3}$$

where, $\mathcal{L}$ is English Language.

Note that, $argmax_x f(x)$ means, "the x such that $f(x)$ is maximum." The equation 7.3 is guaranteed to give optimal sentence W. For a given sentence W and acoustic sequence O we need to compute $P(W|O)$. As per Bayes' theorem,

$$P(x|y) = \frac{P(y|x) * P(x)}{P(y)} \tag{7.4}$$

where, $P(x|y)$ is the probability of occurrence of the event x, given that the event y has already occurred. $P(x)$ is called prior probability.

Accordingly, the equation 7.3 can be expressed as:

$$\hat{W} = argmax_{w \in \mathcal{L}} \frac{P(O|W) * P(W)}{P(O)} \tag{7.5}$$

The prior probability $P(W)$ is estimated by n-gram language model. We can ignore the probability $P(O)$, because it is common for all the sentences. Hence, what we need to compute, reduces to simply:

$$\hat{W} = argmax_{w \in \mathcal{L}} P(O|W) * P(W) \tag{7.6}$$

where, W is most probable sentence, $P(W)$ is prior probability (it is called the *language model*), and $P(O|W)$ is observation likelihood.

We will compute the acoustic model $P(O|W)$ in two steps: first we make assumption that input sequence is a sequence of phones F rather than sequence of acoustic observations. We will show that probability of these phone automata are special case of HMM, and we will show how to extend these models to give probability of a phone sequence given the entire sentence.

There are two algorithms, that simultaneously compute the likelihood of an observation sequence given each sentence, and give us the most likely sentence. These are *viterbi* and *A** algorithms. Finally, we will introduce the HMM approach. These steps are shown in figure 7.2.

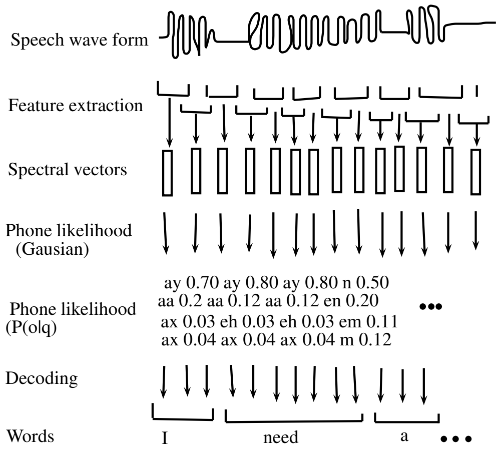

Figure 7.2: Schematic structure of a Speech Recognizer.

References

[1] D. Jurafsky and J. Martin, "Speech and Language Processing," *Pearson India*, 2002, chapter 7.