1.2 A Mathematical Formulation

To discuss the problem of speech recognizer design, we need its mathematical formulation. A precise statement of the problem leads directly to a fruitful decomposition into easier to treat subproblems. Our approach is statistical,[3] so the formulation will involve probabilities. Here it is [3] [4]:

Let **A** denote the acoustic evidence (data) on the basis of which the recognizer will make its decision about which words were spoken. Because we are dealing with digital computers, then without loss of generality we may assume that **A** is a sequence of symbols taken from some (possible very large) alphabet $\mathscr{A}$:

$$\mathbf{A} = a_1, a_2, \ldots, a_m \qquad a_i \in \mathscr{A} \tag{1}$$

The symbols a_i can be thought of as having been generated in time, as indicated by the index i.

Let

$$\mathbf{W} = w_1, w_2, \ldots, w_n \qquad w_i \in \mathscr{V} \tag{2}$$

denote a string of n words, each belonging to a fixed and known vocabulary $\mathscr{V}$.

If $P(\mathbf{W}|\mathbf{A})$ denotes the probability that the words **W** were spoken, given that the evidence **A** was observed, then the recognizer should decide in favor of a word string $\hat{\mathbf{W}}$ satisfying

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{A}) \tag{3}$$

That is, the recognizer will pick the most likely word string given the observed acoustic evidence.

Of course, underlying the target formula (3) is the tacit assumption that all words of a message are equally important to the user, that is, that misrecognition does not carry a different penalty depending on which word was misrecognized. Under this philosophy the warning "*Fire!*" carries no more importance in a crowded theater than an innocuous commercial announcement. But let that possible criticism pass.[4]

3. No advanced results of probability or statistical theory will be used in this self-contained text. The student is required simply to be comfortable with statistical concepts and be able to manipulate them intuitively. So although nothing in this text presumes more than the knowledge of the first four chapters of a book like [5], the required sophistication can probably be gained only by completing an entire course.

The well known Bayes' formula of probability theory allows us to rewrite the right-hand side probability of (3) as

$$P(\mathbf{W}|\mathbf{A}) = \frac{P(\mathbf{W})P(\mathbf{A}|\mathbf{W})}{P(\mathbf{A})} \tag{4}$$

where $P(\mathbf{W})$ is the probability that the word string **W** will be uttered, $P(\mathbf{A}|\mathbf{W})$ is the probability that when the speaker says **W** the acoustic evidence **A** will be observed, and $P(\mathbf{A})$ is the average probability that **A** will be observed. That is,

$$P(\mathbf{A}) = \sum_{\mathbf{W}'} P(\mathbf{W}')P(\mathbf{A}|\mathbf{W}') \tag{5}$$

Since the maximization in (3) is carried out with the variable **A** fixed (there is no other acoustic data save the one we are given), it follows from (3) and (4) that the recognizer's aim is to find the word string $\hat{\mathbf{W}}$ that maximizes the product $P(\mathbf{W})P(\mathbf{A}|\mathbf{W})$, that is, it satisfies

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W})P(\mathbf{A}|\mathbf{W}) \tag{6}$$

1.3 Components of a Speech Recognizer

Formula (6) determines what processes and components are of concern in the design of a speech recognizer.

1.3.1 Acoustic Processing

First, it is necessary to decide what acoustic data **A** will be observed. That is, one needs to decide on a *front end* that will transform the pressure waveform (which is what sound is) into the symbols a_i with which the recognizer will deal. So in principle, this front end includes a microphone whose output is an electric signal, a means of sampling that signal, and a manner of processing the resulting sequence of samples.

4. Strictly speaking, formula (3) is appropriate only if we are after a perfect transcription of the utterance, that is, if one error is as bad as many. Were we to accept that errors are inevitable (which they certainly are) and aim explicitly at minimizing their number, a much more complex formula would be required. So our formula only approximates (it turns out very fruitfully) what we are intuitively after.