Table I. Typical lexicon, with the word *the* having two pronunciations.

Word	Phonetic representation
The	DH AH
The	DH IY
Cat	K AE T
Pig	P IH G
Two	T UW

HYPOTHESIS SEARCH

Three basic components comprise the hypothesis search: a lexicon, a language model, and an acoustic model.

Lexicon

The typical lexicon shown in Table 1 lists each word's possible pronunciations, constructed from phonemes, of which English uses approximately 50. An individual word can have multiple pronunciations, however, which complicates recognition tasks. The system chooses the lexicon on a task-dependent basis, trading off vocabulary size with word coverage. Although a search can easily find phonetic representations for commonly used words in various sources, task-dependent jargon often requires writing out pronunciations by hand.

Language model

The search for the most likely word sequence in Equation 1 requires the computation of two terms, $p(y_1^T|w_1^N)$ and $p(w_1^N)$. The second of these computations is called the *language model*. Its function is to assign a probability to a sequence of words w_1^N.

The simplest way to determine such a probability would be to compute the relative frequencies of different word sequences. However, the number of different sequences grows exponentially with the length of the sequence, making this approach infeasible.

A typical approximation assumes that the probability of the current word depends on the previous two words only, so that the computation can approximate the probability of the word sequence as:

$$p\left(w_1^N\right) \approx p(w_1)p\left(w_2|w_1\right)\prod_{i=3}^{i=N} p\left(w_i|w_{i-1},w_{i-2}\right) \quad (2)$$

The computation can estimate $p(w_i|w_{i-1}, w_{i-2})$ by counting the relative frequencies of word trigrams, or triplets:

$$p\left(w_i|w_{i-1},w_{i-2}\right) \approx N\left(w_i,w_{i-1},w_{i-2}\right)/ N\left(w_{i-1},w_{i-2}\right) \quad (3)$$

where N refers to the associated event's relative frequency. Typically, training such a language model requires using hundreds of millions of words to estimate $p(w_i|w_{i-1}, w_{i-2})$. Even then, many trigrams do not occur in the training text, so the computation must smooth the probability estimates to avoid zeros in the probability assignment.[3]

Acoustic models

An acoustic model computes the probability of feature vector sequences under the assumption that a particular word sequence produced the vectors.

Given speech's inherently stochastic nature, speakers usually do not utter a word the same way twice. The variation in a word's or phoneme's pronunciation manifests itself in two ways: duration and spectral content, also known as acoustic observations. Further, phonemes in the surrounding context can cause variations in a particular phoneme's spectral content, a phenomenon called *coarticulation*.

Hidden Markov models. A *hidden Markov model* offers a natural choice for modeling speech's stochastic aspects. HMMs function as probabilistic finite state machines: The model consists of a set of states, and its topology specifies the allowed transitions between them. At every time frame, an HMM makes a probabilistic transition from one state to another and emits a feature vector with each transition.

Figure 2 shows an HMM for a phoneme. A set of *state transition probabilities*—$p1$, $p2$, and $p3$—governs the possible transitions between states. They specify the probability of going from one state at time t to another state at time $t + 1$. The feature vectors emitted while making a particular transition represent the spectral characteristics of the speech at that point, which vary corresponding to different pronunciations of the phoneme. A *probability distribution* or *probability density function* models this variation. The functions—$p(y|1)$, $p(y|2)$, and $p(y|3)$—could be different for different transitions. Typically, these distributions are modeled as parametric distributions—a mixture of multidimensional gaussians, for example.

The HMM shown in Figure 2 consists of three states. The phoneme's pronunciation corresponds to starting from the first state and making a sequence of transitions to eventually arrive at the third state. The duration of the phoneme equals the number of time frames required to complete the transition sequence. The three transition probabilities implicitly specify a probability distribution that governs this duration. If any of these transitions exhibits high self-loop probabilities, the