model spends more time in the same state, consequently taking longer to go from the first to the third state. The probability density functions associated with the three transitions govern the sequence of output feature vectors.

A fundamental operation is the computation of the likelihood that an HMM produces a given sequence of acoustic feature vectors. For example, assume that the system extracted T feature vectors from speech corresponding to the pronunciation of a single phoneme, and that the system seeks to infer which phoneme from a set of 50 was spoken. The procedure for inferring the phoneme assumes that the ith phoneme was spoken and finds the likelihood that the HMM for this phoneme produced the observed feature vectors. The system then hypothesizes that the spoken phoneme model is the one with the highest likelihood of matching the observed sequence of feature vectors.

If we know the sequence of HMM states, we can easily compute the probability of a sequence of feature vectors. In this case, the system computes the likelihood of the tth feature vector, y_t, using the probability density function for the HMM state at time t. The likelihood of the complete set of T feature vectors is the product of all these individual likelihoods. However, because we generally do not know the actual sequence of transitions, the likelihood computation process sums all possible state sequences. Given that all HMM dependencies are local, we can derive efficient formulas for performing these calculations recursively.[3]

Parameter estimation. The state transition probabilities and the gaussian means and variances that model the feature vectors' probability density functions parameterize the HMM's different states. Before using an HMM to compute the likelihood values of feature vector sequences, we must train the HMMs to estimate the model's parameters. This process assumes the availability of a large amount of training data, which consists of the spoken word sequence and the feature vectors extracted from the speech signal.

Researchers commonly use the maximum likelihood estimation process training paradigm for this task. Given that we know the correct word sequence corresponding to the feature vector sequence, the ML estimation process tries to choose the HMM parameters that maximize the training feature vectors' likelihood, computed using the HMM for the correct word sequence. If y_1^T represents the stream of T acoustic observations, and w_1^N represents the correct word sequence, the ML estimate for the parameter θ is

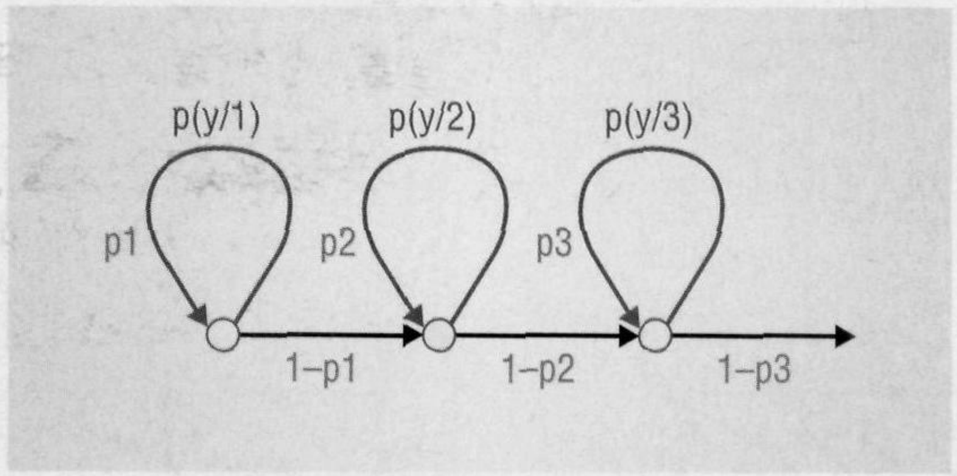

Figure 2. Hidden Markov model for a phoneme. State transition probabilities p1, p2, and p3 govern the possible transitions between states.

$$\hat{\theta}_{ML} = \arg\max_{\theta} \log\left[p_\theta\left(y_1^T \middle| w_1^N\right)\right] \quad (4)$$

The system begins the training process by constructing an HMM for the correct word sequence. First, it constructs the HMMs for each word by concatenating the HMMs for the phonemes that comprise the word. Subsequently, it concatenates the word HMMs to form the HMM for the complete utterance. For example, the HMM for the utterance "We were" would be the concatenation of the HMMs for the four phonemes "W IY W ER."

The training process assumes that the system can generate the acoustic observations y_1^T by traversing the HMM from its initial state to its final state in T time frames. However, because the system cannot trace the actual state sequence, the ML estimation process assumes that this state sequence is hidden and averages all possible state sequence values.

By using s_t to denote the hidden state at time t, then making various assumptions, the system can express the maximization of Equation 4 in terms of the HMM's hidden states, as follows:

$$\arg\max_{\hat{\theta}} \sum_{t=1}^{T} \sum_{s_t} p_\theta\left(s_t \middle| y_1^T\right) \log\left[p_{\hat{\theta}}\left(y_t \middle| s_t\right)\right] \quad (5)$$

The system uses an iterative process to solve Equation 5, with each iteration involving an expectation step and a maximization step.[3] The first step involves the computation of $p_\theta(s_t|y_1^T)$, which is the posterior probability—or count of a state—conditioned on all the acoustic observations. The system uses the current HMM parameter estimates and the Forward-Backward algorithm[3] to perform this computation. The second step involves choosing the parameter $\hat{\theta}$ to maximize Equation 5. When the probability density functions are gaussians, the computation can derive closed-form expressions for this step.

Coarticulation. So far, we have assumed that the fundamental acoustic units are phonemes and that the system uses HMMs to model the duration and acoustic variation associated with the phonemes' pronunciation. However, in some cases, the phonemes in the surrounding context affect a particular phoneme's acoustic variation. This coarticulation phenomenon is particularly prevalent in spontaneous speech, which the speaker does not enunciate carefully. The system models coarticulation by assuming that the density of the obser-