Hidden Markov Models

Hermann Ney

Final version

HIDDEN MARKOV MODELS (HMMs) are stochastic finite-state automata (or, equivalently, stochastic regular grammars) with either continuous-valued or discrete-valued observations. HMMs have found widespread use in automatic speech recognition, where they provide an efficient mechanism for modelling speaking rate variations within a probabilistic framework. An HMM computes the probability (density) $Pr(x_1...x_T|w_1...w_N)$ of the acoustic–phonetic model, which is the conditional probability of observing the acoustic vectors $x_1...x_T$ over time $t = 1, ..., T$ when the speaker utters the words $w_1...w_N$. An acoustic vector x_t is computed typically every 10 msec and represents the short-term magnitude spectrum of the acoustic signal.

For a large–vocabulary system, there is typically a set of basic recognition units that are smaller than whole words. Examples of these so–called subword units are phonemes, demisyllables or syllables. The word models are then obtained by concatenating the subword models according to the phonetic transcription of the words in a pronunciation dictionary. In virtually all systems, these subword units are modeled by HMMs.

Although the theory of HMMs applies to *any* structure of the finite-state automaton, a linear arrangement of states is widely used. The states can be interpreted as time points on a normalized time axis. To allow speaking rate variations, there are typically three types of possible transitions by which each state can be left: move to the next state, loop back to the same state and skip to the next but one state. Such a structure is depicted in Fig. 1 along with the resulting trellis that shows the possible time alignment paths.

To obtain a quantitative description of an HMM, we consider the case of whole-word models. For a given state s' in a word model w, we have a transition probability $p(s|s', w)$ for going to state s. In addition, there is an emission probability (density) $p(x_t|s, w)$ for observing vector x_t at time

t when reaching state s in word model w. Typically, the product of the emission and transition probabilities is used:

$$p(x_t, s|s'; w) = p(s|s'; w) \cdot p(x_t|s; w) \ ,$$

which is the conditional probability that, given state s' in word model w, the acoustic vector x_t is observed and the state s is reached.

Since the states are hidden, i.e. abstractions of the model used, and cannot be observed directly, it is – from a strict statistical point of view – necessary to sum over all possible state sequences $s_1^T = s_1 ... s_t ... s_T$ to compute the probability $Pr(x_1 ... x_T|w)$:

$$\begin{aligned} Pr(x_1 ... x_T|w) &= \sum_{\{s_1^T\}} \prod_{t=1}^{T} [p(x_t, s_t|s_{t-1}; w)] \\ &\cong \max_{\{s_1^T\}} \prod_{t=1}^{T} [p(x_t, s_t|s_{t-1}; w)] \end{aligned}$$

where we have replaced the sum by the maximum. This so-called maximum or Viterbi approximation is found to be sufficient in most practical applications. Both the sum and the maximum of all state sequences can be computed efficiently by exploiting the first-order dependency structure of the HMM (forward recursion for the sum, dynamic programming for the maximum).

To estimate the free parameters of an HMM from training data, powerful algorithms like the expectation-maximization algorithm are available, see Jelinek (1998), Rabiner and Juang (1993). In particular, the training can be performed in such a way that no manual segmentation of the speech signal into words or phones is required.

The HMM approach fits directly into Bayes decision rule for automatic speech recognition and allows the interdependence of several operations to be handled using a single consistent criterion: identification of spoken words (and phones), nonlinear time alignment, (implicit) segmentation of the speech signal into phones and words, and taking into account the language model. This approach results in a huge search space, which, however, can be handled efficiently by dynamic programming beam search, see Ney and Ortmanns (2000).

In addition to speech recognition, HMMs or related approaches are used successfully also in other linguistic applications, including dynamic → *Optical*

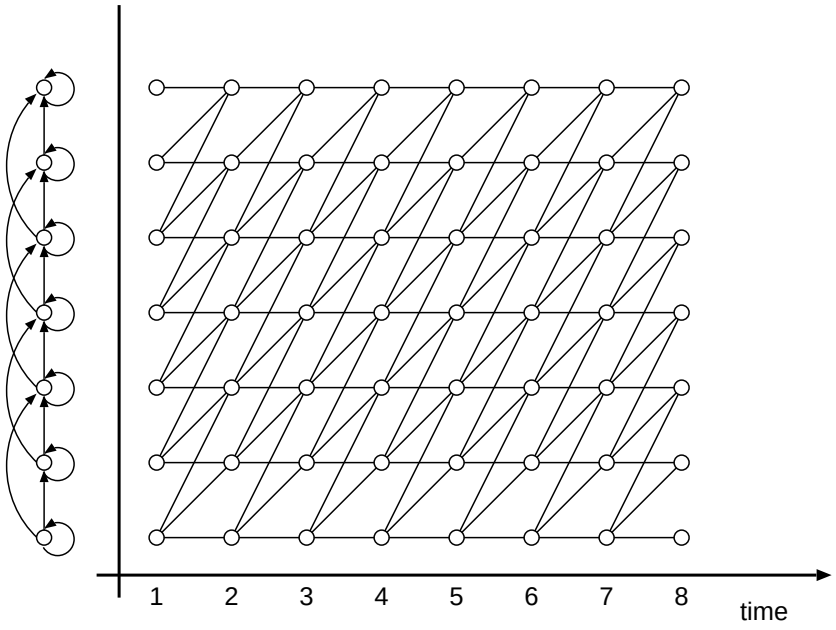
Character Recognition, statistical → *Machine Translation* of (written and spoken) language, statistical → *Grammatical Tagging*, and probabilistic → *Context Free Grammars*.

References F. Jelinek: *Speech Recognition by Statistical Methods*. MIT Press, Cambridge, MA,1998.

H. Ney, S. Ortmanns: Progress in Dynamic Programming Search for LVCSR. *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1224-1240, Aug. 2000.

L. R. Rabiner, B. H. Juang: *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

Index Acoustic-Phonetic Modelling, Phoneme and Word Models, Speech Recognition, Maximum Likelihood Training, Dynamic Programming, Beam Search, Nonlinear Time Alignment, Interdependence of Operations in Speech Recognition.


1
2
3
4
5
6
7
8
time