DOWNLOAD PDF HIDDEN MARKOV MODEL SPEECH RECOGNITION

Chapter 1 : Hidden Markov model - Wikipedia

Foundations and TrendsR in Signal Processing Vol. 1, No. 3 () c M. Gales and S. Young DOI: / The Application of Hidden Markov Models in Speech Recognition.

This page is under construction. Find Pr sigma lambda: Find the most likely state trajectory given the model and observations. Motivation A discrete-time, discrete-space dynamical system governed by a Markov chain emits a sequence of observable outputs: From the observable sequence of outputs, infer the most likely dynamical system. The result is a model for the underlying process. Alternatively, given a sequence of outputs, infer the most likely sequence of states. We might also use the model to predict the next observation or more generally a continuation of the sequence of observations. Hidden Markov models are used in speech recognition. Suppose that we have a set W of words and a separate training set for each word. Build an HMM for each word using the associated training set. When presented with a sequence of observations sigma, choose the word with the most likely model, i. Compute the forward alpha values: Computing the backward beta values: Baum-Welch Algorithm Intuition To solve Problem 3 we need a method of adjusting the lambda parameters to maximize the likelihood of the training set. To estimate the lambda parameters for this Markov chain it is enough just to calculate the appropriate frequencies from the observed sequence of outputs. These frequencies constitute sufficient statistics for the underlying distributions. In the hidden case, we use expectation maximization EM as described in [Dempster et al. Instead of calculating the required frequencies directly from the observed outputs, we iteratively estimated the parameters. We start by choosing arbitrary values for the parameters just make sure that the values satisfy the requirements for probability distributions. We then compute the expected frequencies given the model and the observations. The expected frequencies are obtained by weighting the observed transitions by the probabilities specified in the current model. The expected frequencies so obtained are then substituted for the old parameters and we iterate until there is no improvement. On each iteration we improve the probability of O being observed from the model until some limiting probability is reached. This iterative procedure is guaranteed to converge on a local maximum of the cross entropy Kullback-Leibler performance measure. One advantage of this approach is that it extends easily to the case in which the hidden part of the model is factored into some number of state variables. This addition results in the network shown in Figure 2. The computational picture is more complicated and depends on the specifics of the update algorithm. It is important to point out, however, that there is a wide range of update algorithms, both approximate and exact, to choose from. See [Charniak,] for applications in natural language processing including part of speech tagging. Charniak [] provides lots of examples that provide useful insight. Rabiner and Juang [] also discuss variant algorithms for continuous observation spaces using multivariate Gaussian models.

Chapter 2 : Hidden Markov Models

2 Acknowledgements Much of this talk is derived from the paper "An Introduction to Hidden Markov Models", by Rabiner and Juang and from the talk "Hidden Markov Models: Continuous Speech.

Early work[edit] In three Bell Labs researchers, Stephen. Their system worked by locating the formants in the power spectrum of each utterance. Gunnar Fant developed the source-filter model of speech production and published it in , which proved to be a useful model of speech production. Raj Reddy was the first person to take on continuous speech recognition as a graduate student at Stanford University in the late s. Previous systems required the users to make a pause after each word. Also around this time Soviet researchers invented the dynamic time warping DTW algorithm and used it to create a recognizer capable of operating on a word vocabulary. Although DTW would be superseded by later algorithms, the technique of dividing the signal into frames would carry on. Achieving speaker independence was a major unsolved goal of researchers during this time period. In , DARPA funded five years of speech recognition research through its Speech Understanding Research program with ambitious end goals including a minimum vocabulary size of 1, words. It was thought that speech understanding would be key to making progress in speech recognition, although that later proved to not be true. Four years later, the first ICASSP was held in Philadelphia , which since then has been a major venue for the publication of research on speech recognition. Katz introduced the back-off model in , which allowed language models to use multiple length n-grams. As the technology advanced and computers got faster, researchers began tackling harder problems such as larger vocabularies, speaker independence, noisy environments and conversational speech. In particular, this shifting to more difficult tasks has characterized DARPA funding of speech recognition since the s. For example, progress was made on speaker independence first by training on a larger variety of speakers and then later by doing explicit speaker adaptation during decoding. Further reductions in word error rate came as researchers shifted acoustic models to be discriminative instead of using maximum likelihood estimation. This processor was extremely complex for that time, since it carried However, nowadays the need of specific microprocessor aimed to speech recognition tasks is still alive: By this point, the vocabulary of the typical commercial speech recognition system was larger than the average human vocabulary. Handling continuous speech with a large vocabulary was a major milestone in the history of speech recognition. Huang went on to found the speech recognition group at Microsoft in Apple originally licensed software from Nuance to provide speech recognition capability to its digital assistant Siri. Four teams participated in the EARS program: EARS funded the collection of the Switchboard telephone speech corpus containing hours of recorded conversations from over speakers. The recordings from GOOG produced valuable data that helped Google improve their recognition systems. Google voice search is now supported in over 30 languages. In the United States, the National Security Agency has made use of a type of speech recognition for keyword spotting since at least Recordings can be indexed and analysts can run queries over the database to find conversations of interest. Some government research programs focused on intelligence applications of speech recognition, e. Voice recognition[edit] What, by early s was often called speech recognition, so as to differentiate from speaker recognition, was also called voice recognition; this is what was commonly used. A ad for a doll carried the tagline "Finally, the doll that understands you. Researchers have begun to use deep learning techniques for language modeling as well. In the long history of speech recognition, both shallow form and deep form e. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning starting around â€” that had overcome all these difficulties. Hidden Markov models HMMs are widely used in many systems. Language modeling is also used in many other natural language processing applications such as document classification or statistical machine translation. Hidden Markov models[edit] Main article: Hidden Markov model Modern general-purpose speech recognition systems are based on Hidden Markov Models. These are statistical models that output a sequence of symbols or quantities. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary

signal. In a short time-scale e. Speech can be thought of as a Markov model for many stochastic purposes. Another reason why HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. In speech recognition, the hidden Markov model would output a sequence of n-dimensional real-valued vectors with n being a small integer, such as 10 , outputting one of these every 10 milliseconds. The vectors would consist of cepstral coefficients, which are obtained by taking a Fourier transform of a short time window of speech and decorrelating the spectrum using a cosine transform , then taking the first most significant coefficients. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians, which will give a likelihood for each observed vector. Each word, or for more general speech recognition systems , each phoneme , will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes. Described above are the core elements of the most common, HMM-based approach to speech recognition. Modern speech recognition systems use various combinations of a number of standard techniques in order to improve results over the basic approach described above. A typical large-vocabulary system would need context dependency for the phonemes so phonemes with different left and right context have different realizations as HMM states ; it would use cepstral normalization to normalize for different speaker and recording conditions; for further speaker normalization it might use vocal tract length normalization VTLN for male-female normalization and maximum likelihood linear regression MLLR for more general speaker adaptation. The features would have so-called delta and delta-delta coefficients to capture speech dynamics and in addition might use heteroscedastic linear discriminant analysis HLDA ; or might skip the delta and delta-delta coefficients and use splicing and an LDA -based projection followed perhaps by heteroscedastic linear discriminant analysis or a global semi-tied co variance transform also known as maximum likelihood linear transform , or MLLT. Many systems use so-called discriminative training techniques that dispense with a purely statistical approach to HMM parameter estimation and instead optimize some classification-related measure of the training data. Decoding of the speech the term for what happens when the system is presented with a new utterance and must compute the most likely source sentence would probably use the Viterbi algorithm to find the best path, and here there is a choice between dynamically creating a combination hidden Markov model, which includes both the acoustic and language model information, and combining it statically beforehand the finite state transducer , or FST, approach. A possible improvement to decoding is to keep a set of good candidates instead of just keeping the best candidate, and to use a better scoring function re scoring to rate these good candidates so that we may pick the best one according to this refined score. The set of candidates can be kept either as a list the N-best list approach or as a subset of the models a lattice. Re scoring is usually done by trying to minimize the Bayes risk [62] or an approximation thereof: Instead of taking the source sentence with maximal probability, we try to take the sentence that minimizes the expectancy of a given loss function with regards to all possible transcriptions i. The loss function is usually the Levenshtein distance , though it can be different distances for specific tasks; the set of possible transcriptions is, of course, pruned to maintain tractability. Efficient algorithms have been devised to re score lattices represented as weighted finite state transducers with edit distances represented themselves as a finite state transducer verifying certain assumptions. Dynamic time warping Dynamic time warping is an approach that was historically used for speech recognition but has now largely been displaced by the more successful HMM-based approach. Dynamic time warping is an algorithm for measuring similarity between two sequences that may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video the person was walking slowly and if in another he or she were walking more quickly, or even if there were accelerations and deceleration during the course of one observation. A well-known application has been automatic speech recognition, to cope with different speaking speeds. In general, it is a method that allows a computer to find an optimal match between two given sequences e. That is, the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of hidden Markov models. Artificial neural network Neural networks emerged as an attractive acoustic modeling approach in ASR in the late s. Since then, neural networks have been used in many aspects of speech

recognition such as phoneme classification, [64] isolated word recognition, [65] audiovisual speech recognition, audiovisual speaker recognition and speaker adaptation. In contrast to HMMs, neural networks make no assumptions about feature statistical properties and have several qualities making them attractive recognition models for speech recognition. When used to estimate the probabilities of a speech feature segment, neural networks allow discriminative training in a natural and efficient manner. Few assumptions on the statistics of input features are made with neural networks. However, in spite of their effectiveness in classifying short-time units such as individual phonemes and isolated words, [66] neural networks are rarely successful for continuous recognition tasks, largely because of their lack of ability to model temporal dependencies. Deep Neural Networks and Denoising Autoencoders [70] were also being experimented with to tackle this problem in an effective manner. Due to the inability of feedforward Neural Networks to model temporal dependencies, an alternative approach is to use neural networks as a pre-processing e. Deep feedforward and recurrent neural networks[edit] Main article: Deep learning A deep feedforward neural network DNN is an artificial neural network with multiple hidden layers of units between the input and output layers. DNN architectures generate compositional models, where extra layers enable composition of features from lower layers, giving a huge learning capacity and thus the potential of modeling complex patterns of speech data. This principle was first explored successfully in the architecture of deep autoencoder on the "raw" spectrogram or linear filter-bank features, [79] showing its superiority over the Mel-Cepstral features which contain a few stages of fixed transformation from spectrograms. The true "raw" features of speech, waveforms, have more recently been shown to produce excellent larger-scale speech recognition results. End-to-end models jointly learn all the components of the speech recognizer. This is valuable since it simplifies the training process and deployment process. For example, a n-gram language model is required for all HMM-based systems, and a typical n-gram language model often takes several gigabytes in memory making them impractical to deploy on mobile devices. Jointly, the RNN-CTC model learns the pronunciation and acoustic model together, however it is incapable of learning the language due to conditional independence assumptions similar to a HMM. Consequently, CTC models can directly learn to map speech acoustics to English characters, but the models make many common spelling mistakes and must rely on a separate language model to clean up the transcripts. Later, Baidu expanded on the work with extremely large datasets and demonstrated some commercial success in Chinese Mandarin and English. Attention-based ASR models were introduced simultaneously by Chan et al. Unlike CTC-based models, attention-based models do not have conditional-independence assumptions and can learn all the components of a speech recognizer including the pronunciation, acoustic and language model directly. This means, during deployment, there is no need to carry around a language model making it very practical for deployment onto applications with limited memory. By the end of , the attention-based models have seen considerable success including outperforming the CTC models with or without an external language model. Following the audio prompt, the system has a "listening window" during which it may accept a speech input for recognition. Voice recognition capabilities vary between car make and model. Some of the most recent[when? With such systems there is, therefore, no need for the user to memorize a set of fixed command words. Front-end speech recognition is where the provider dictates into a speech-recognition engine, the recognized words are displayed as they are spoken, and the dictator is responsible for editing and signing off on the document. Back-end or deferred speech recognition is where the provider dictates into a digital dictation system, the voice is routed through a speech-recognition machine and the recognized draft document is routed along with the original voice file to the editor, where the draft is edited and report finalized. Deferred speech recognition is widely used in the industry currently. One of the major issues relating to the use of speech recognition in healthcare is that the American Recovery and Reinvestment Act of ARRA provides for substantial financial benefits to physicians who utilize an EMR according to "Meaningful Use" standards. A more significant issue is that most EHRs have not been expressly tailored to take advantage of voice-recognition capabilities. By contrast, many highly customized systems for radiology or pathology dictation implement voice "macros", where the use of certain phrases â€” e.

Chapter 3 : HTK Speech Recognition Toolkit

type of model is Gaussian Model, Poisson Model, Markov Model and Hidden Markov model. Speech Recognition: Speech recognition is a process of converting speech signal to a se-.

From the perspective described above, this can be thought of as the probability distribution over hidden states for a point in time k in the past, relative to time t. The forward-backward algorithm is an efficient method for computing the smoothed values for all hidden state variables. Most likely explanation[edit] The task, unlike the previous two, asks about the joint probability of the entire sequence of hidden states that generated a particular sequence of observations see illustration on the right. An example is part-of-speech tagging , where the hidden states represent the underlying parts of speech corresponding to an observed sequence of words. In this case, what is of interest is the entire sequence of parts of speech, rather than simply the part of speech for a single word, as filtering or smoothing would compute. This task requires finding a maximum over all possible state sequences, and can be solved efficiently by the Viterbi algorithm. Statistical significance[edit] For some of the above problems, it may also be interesting to ask about statistical significance. What is the probability that a sequence drawn from some null distribution will have an HMM probability in the case of the forward algorithm or a maximum state sequence probability in the case of the Viterbi algorithm at least as large as that of a particular output sequence? A concrete example[edit] Consider two friends, Alice and Bob, who live far apart from each other and who talk together daily over the telephone about what they did that day. Bob is only interested in three activities: The choice of what to do is determined exclusively by the weather on a given day. Alice has no definite information about the weather, but she knows general trends. Based on what Bob tells her he did each day, Alice tries to guess what the weather must have been like. Alice believes that the weather operates as a discrete Markov chain. There are two states, "Rainy" and "Sunny", but she cannot observe them directly, that is, they are hidden from her. On each day, there is a certain chance that Bob will perform one of the following activities, depending on the weather: Since Bob tells Alice about his activities, those are the observations. The entire system is that of a hidden Markov model HMM. Alice knows the general weather trends in the area, and what Bob likes to do on average. In other words, the parameters of the HMM are known. They can be represented as follows in Python: A similar example is further elaborated in the Viterbi algorithm page. Learning[edit] The parameter learning task in HMMs is to find, given an output sequence or a set of such sequences, the best set of state transition and emission probabilities. The task is usually to derive the maximum likelihood estimate of the parameters of the HMM given the set of output sequences. No tractable algorithm is known for solving this problem exactly, but a local maximum likelihood can be derived efficiently using the Baumâ€”Welch algorithm or the Baldiâ€—Chauvin algorithm. The Baumâ€—Welch algorithm is a special case of the expectation-maximization algorithm. If the HMMs are used for time series prediction, more sophisticated Bayesian inference methods, like Markov chain Monte Carlo MCMC sampling are proven to be favorable over finding a single maximum likelihood model both in terms of accuracy and stability.

Chapter 4 : Speech recognition - Wikipedia

Speech recognition using hidden Markov model The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice.

You can also check a textbook like Spoken Language Processing which gives a good description for the subject. The sequence of observable events in speech recognition is sequence of audio frames. Each frame is roughly 20ms of sound. The sequence of unobservable events is roughly the sequence of phonemes. Actually it is more complex, but you can think phonemes. Beside HMM model which is just mathematical object, there is an important part about decoding algorithm called Viterbi search which finds the best match between observable and hidden states according to the probabilities. This algorithm efficiently evaluates all possible breakdowns and find the best one. That best one would be the decoding result. There is no such thing as "hmm in state k". We consider frame 1 and say it corresponds to "k", then we consider frame 2 and decide does it correspond to "k", to "a" or to "e". For that we use previous state for previous frame and also acoustic match between the frame 2 and all three states. This acoustic match is usually estimated with separate model, for example gaussian mixture model, do not confuse it with hidden markov model. Both models are estimated from the corpus. After we store some possible decisions for frame 2 we move to frame 3 to decide does it belong to any of expected hidden states. Notice we do not keep 1 best decision but multiple possible decisions on the way because locally best decision 2 corresponds to a might not be globally best 2 corresponds to e. In the end of decoding we have a full relation between hidden and observable states and we can estimate the probability of this relation using HMM. It compares probabilities of breakdowns combined with the GMM score for this frame to update probabilities of breakdowns including this new frame. GMM score tells how good is match of the audio to the expected sound of "a" and is trained from the database. And it says that the phonemes are the hidden parts And we can only see, which phonemes it chose, after it processed the entire word and the outcome is either correct or not? We can only see phonemes after processing the entire word. Locally you can not guarantee that, you need to compare global picture or at least to do some iterations after current phoneme. Thats why you have to keep multiple decoding results during search, not just the best single one.

Chapter 5 : speech recognition - how do hidden markov models recognize a word? - Stack Overflow

Continuous Speech Recognition Using Hidden Markov Models Joseph Picone Stochastic signal processing techniques have pro- foundly changed our perspective on speech processing.

Sponsors What is HTK? HTK is primarily used for speech recognition research although it has been used for numerous other applications including research into speech synthesis, character recognition and DNA sequencing. HTK is in use at hundreds of sites worldwide. HTK consists of a set of library modules and tools available in C source form. The tools provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. The software supports HMMs using both continuous density mixture Gaussians and discrete distributions and can be used to build complex HMM systems. The HTK release contains extensive documentation and examples. In Entropic Research Laboratory Inc. See History of HTK for more details. While Microsoft retains the copyright to the original HTK code, everybody is encouraged to make changes to the source code and contribute them for inclusion in HTK3. You must then register for a username and password which will allow you to download the HTK Book and source code. Registration is free but does require a valid e-mail address; your password for site access will be sent to this address. Thanks for the feedback and suggestions from various users. We are still working on more substantial updates to HTK 3. This can be downloaded from the HTK downloads page. Note that the samples package is now included with the HTK 3. HDecode is still an additional download due to its separate license. Key features of HTK 3. Only a simple build procedure is included which will require some manual configuration. A more automatic configuration will be available in future as well as support for other platforms. This is an alpha version of the book and so is in some places incomplete. The book also includes extended tutorial information for using the new HTK features, and includes a new section of tutorial examples using the Resource Management task that illustrate new and old functionality. The scripts that are provided for this task may well be of use more generally. In future we intend to both extend the functionality of HTK 3. We are currently preparing a new major release, HTK 3. The key features of HTK 3. The meeting will be held from 6pm to 8: We will have a short presentation covering new features in the 3. Some examples of using the HTK large vocabulary decoder and discriminative training tools will also be shown. This is followed by an open discussion and networking. We will also provide some liquid refreshments. Please feel free to forward this announcement to other researchers interested in HTK. We hope to see many of you in Taipei. These meetings are intended to provide a forum for users of HTK and other researchers interested in speech recognition toolkits to exchange ideas and discuss future plans. The meeting will be held from 6pm to 10pm, Thursday, 19th April in the "Ilima" meeting room, Hotel Ala Moana right cross Atkinson drive, opposite to the Hawaii convention center. We will have a short presentation covering: We hope to see many of you in Honolulu. HERest now incorporates the adaptation transform generation that was previously performed in HEAdapt. The range of linear transformations and the ability to combine transforms hierarchically has now been included. A new version 1. For further information see the ATK page.