Speech Recognition: General Idea and Overview

Kewal Mehta[1], Amitesh Dubey[2], Rahul Kalsariya[3], Viraj Prajapati[4], Prof. Suvarna Pansambal[5]

[1,2,3,4]BE, Department of Computer Engineering, Atharva College of Engineering, Mumbai, India
[5]Head of Department, Department of Computer Engineering, Atharva College of Engineering, Mumbai, India

---***---

Abstract - *Voice recognition system is a software that lets the user control computer functions and helps dictate various texts. This paper presents a general idea and an overview of Speech Recognition. We discuss about the general working of speech recognition and some popular algorithms that are used in modern day speech recognition devices. Speech is something that is used everyday and is the most common means of communication between humans, but nowadays humans are not restricted to communicating with just humans, communication of humans with machines is also possible nowadays due to the advancements in the fields of Artificial Intelligence, Machine Learning and Deep learning. This interaction between humans and computers is done using different interfaces, this is termed as human computer interaction(HCI). This paper primarily focuses on basic working and most recognized algorithms of speech recognition which is one of the most important domain in the field of artificial intelligence. The paper also gives detailed knowledge about the various steps of a basic speech recognition system such as pre-processing, feature extraction and reorganization. The paper also gives detailed explanation of various algorithms that are very popular in the field of artificial intelligence such as PLP(Perceptual linear programming), NLP(Natural language processing), DTW(Dynamic time wrap), HMM(Hidden Markov model), N-grams and shows ways to implement these algorithms in the speech recognition devices.*

***Key Words*: Speech Recognition, Artificial Intelligence, Machine Learning, Deep Learning, Human computer Interaction, Perceptual Linear Programming, Natural Language Processing, Dynamic time wrap, Hidden Markov model.**

1. INTRODUCTION

Speech recognition or Speech to text is the ability of a machine or program to identify spoken words on the external side and convert them in the form of text which is readable. Basic Speech recognition software only have access to limited vocabulary, words and phrases and can only identify these speeches only if they spoken clearly. The more sophisticated and technically sound software have the ability to accept and process complex quotes, accents and also languages. Speech recognition incorporates different fields of research in computer science, linguistics and computer engineering. Many modern devices may have speech recognition functions in them to allow for easier or hands - free use of a device. Speech recognition works using algorithms through acoustic and language modeling. So, Speech recognition basically works by breaking down the audio of a speech recording into individual sounds it then analyzes each sound by using algorithms to find the most probable word fit in that language and translating those sounds into text. This is the basic working of speech recognition. However, more advanced speech recognition software make the use AI and ML. These systems will use grammar, structure, syntax as well as composition of audio and voice signals in order to process speech. Software using machine learning will learn more the more it is used, so it may be easier to learn concepts like accents. Speech recognition is one of the leading applications of machine learning.[1]

1.1 Basic Model of Speech Recognition

Speech recognition, also called as Automatic speech recognition (ASR), computer speech recognition or speech-to-text, is a capability which enables a program or software to process human speech into written format. It is basically a method of active communication between a machine and a human. Speech recognition is commonly confused with voice recognition. Speech recognition primarily focuses on converting a human speech into written text whereas, voice recognition focuses on identifying an individual's voice. Many speech recognition devices are available. Speech recognition uses many interdisciplinary technologies ranging from Pattern recognition, Signal processing, Natural language processing implementing to unified statistical framework. They integrate grammar, syntax and composition of audio and voice signals to understand and process human speech. Over time, the software and programs learn the usual behaviors and requirements of the user and evolve after every interaction. This is where machine learning comes into effect. Majority of the good systems allow the companies or the users to customize and adapt the technology as per their requirement. [2]

1.2 Basic Speech Recognition System

The basic speech recognition system goes as follows.

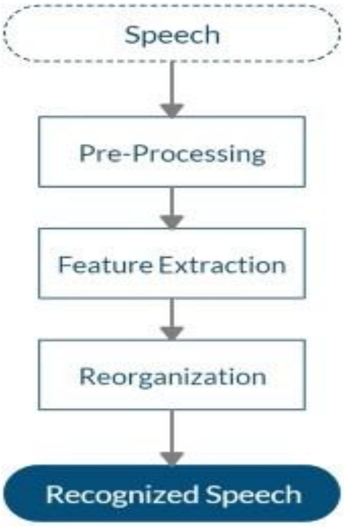

Fig 1:Block Model of Speech Recognition System

(i) ***Speech Capturing***: The capturing of the speech is done using microphones. The analog signal is converted to digital signal using a sound card.

(ii) ***Pre-Processing****:* After the completion of the sound capturing process, the speech or sound is available in the form of continuous wave. The pre-processing section has four stages:

(a) Background noise and silence removal

(b) Pre emphasis filter

(c) Blocking into frames

(d) Windowing

*(iii)****Feature Extraction****:* This is the process of transforming speech into parameters which can represent speech signal speech signal in terms of feature vectors. Using digital filter, Fourier transformation, linear predictive coding, we can extract feature vectors. These speech signals should belong to the same feature vector when changing the speakers. Linear predictive coding is the best method for feature extraction.

(iv)**Reorganization**: Reorganization is divided into two parts:

(a) Training Part: In the training part the system experiences and learns. This means that the system learns and experiences the process and starts updating itself accordingly. The current speech recognition technologies do not allow the real time implementation of models comparable to human complexity.

(b) Testing Part: The testing part is the portion of reorganization where the testing of the implemented actions is done. [5]

2. Algorithms in Speech Recognition

The most common algorithms used in modern day speech recognition systems are:

2.1 Perceptual linear prediction(PLP): PLP models the human speech based on the concept of psychophysics of hearing to derive an estimate of auditory spectrum. PLP discards the irrelevant sections of the speech, thus improving the speech recognition rate. In PLP technique, several known properties of hearing are simulated by practical engineering approximations. The basic block diagram of a PLP system is as given below:

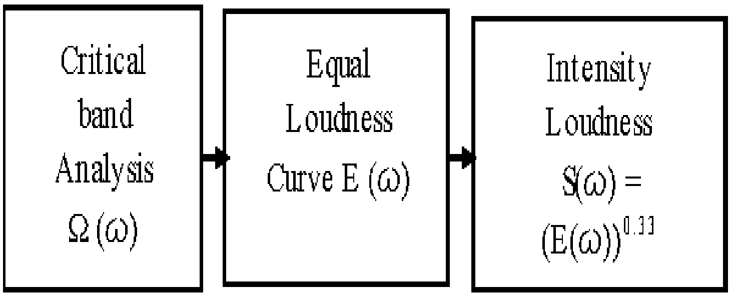
Critical band Analysis Ω (ω)
Equal Loudness Curve E (ω)
Intensity Loudness S(ω) = (E(ω))^0.33

Fig 2: Block diagram of PLP System

The PLP approximates three major perceptual aspects, namely critical band analysis, equal loudness curve and intensity loudness, the auditory spectrum is then approximated by an autoregressive all-pole model. Detailed steps for PLP are shown below.[3]

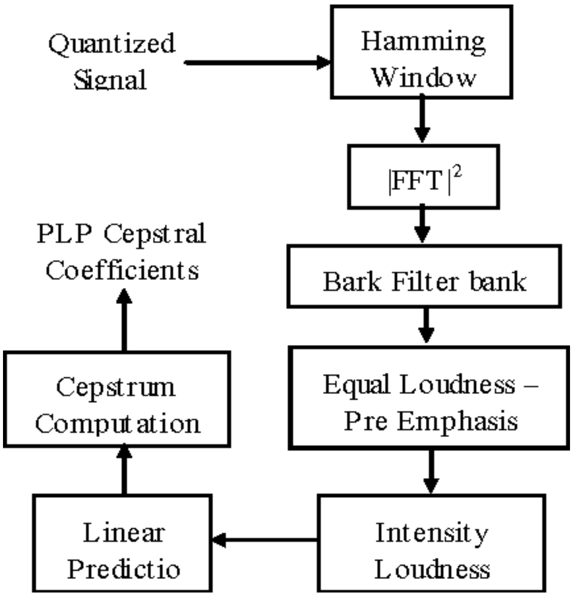

Fig 3: PLP Procedure Block Diagram

Advantages:

1. PLP coefficients are generally used because they approximate the high energy regions of the speech spectrum well. They simultaneously smooth out the fine harmonic structure which is usually the characteristic of a single unit and not the underlying unit.

2. The resolution of the hearing ability of humans is linear up to 800 or 1000 Hz, but decreases with increasing frequency above the linear range.

3. PLP incorporates critical band spectral resolution into its spectrum estimate by remapping the frequency axis to bark scale and integrating the energy in the critical bands to produce a critical-band spectrum approximation. [13]

2.1.1 Natural language processing (NLP): NLP is a subset of artificial intelligence that focuses on the system development which allows systems to communicate with humans in everyday language. In other words, NLP learns how people communicate and teach machines to learn that behaviour. The rise of Machine learning (ML) and innovations in Deep learning (DL) is targeting NLP. Ultimately the goal is to interact with the machines in a natural and a human-like way. The recent advances in NLP is making is possible because of the increasing computing capacity (Moore's law). NLP is now able to process significant amount of data in a very short time frame, owing it to the significant advances in the field of AI, ML, DL. NLP is divided into two categories-

(i) Natural language understanding (NLU)

(ii) Natural language generation (NLG)[5]

NLP is not necessarily an algorithm used for speech recognition specifically, It is the area in the artificial intelligence that focuses on the interaction between machines and humans through speech and text. Hence, it is used in speech recognition devices in the form of conversion stage, that is, it helps in the conversion process of the text to speech.

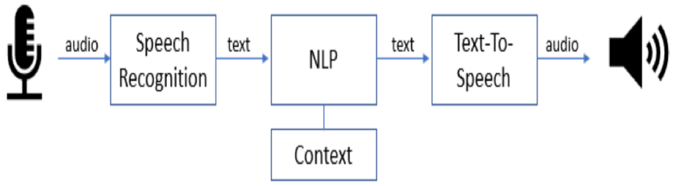

Fig 4:Process Block for NLP

The diagram above shows the basic process of speech recognition using NLP. The audio is sent into a speech recognition algorithm, where it gets converted to text. It is then sent into NLP (Natural Language Processing) and it is then converted into speech to text. In certain cases as shown above, the text can be converted to speech as well. 'Siri' from Apple is the best example of a speech recognition system using NLP.[7]

Advantages:

1. Users get response to their questions on a subject within seconds.

2. NLP system provides answers to questions in natural language.

3. The accuracy of the answers increases with the amount of relevance in the question.

4. NLP provides answers to the point and does not provide unnecessary information.

5. It helps structuring the highly unstructured data source.

Disadvantages:

1. Complex queries: the system will not be able provide the correct answers accurately if the questions stated are poorly framed or ambiguous.

2. The NLP systems are usually built for specific tasks, hence, they are unable to adapt to new domains or problems because of the limited ability.

3. NLP does not have a UI (User interface) which further reduces the possibilities to interact with the system.[14]

2.2 Dynamic Time Wrap (DTW): DTW is a technique that can find an optimal match between two given sequences of speech, this technique supports non-linear mapping from one signal to another by reducing the distance between the two signals. DTW is basically a template based method. It is used to check the compatibility of the sound. DTW is a method to measure the similarity of a pattern which have different time zones. The lesser the distance produced, the more the similarity between the sound patterns. If both the sound patterns are similar then both the voices are said to be the same. The initial data of the speech recognition is produced into frequency waves. Different timing of speech alignment is a core problem for distance measurement in speech recognition. Two words from same words from the same user can have different timings. For example, two can be pronounced as two or twoo. Hence, the time intervals between the two words is different. This issue can be resolved using DTW. [8]

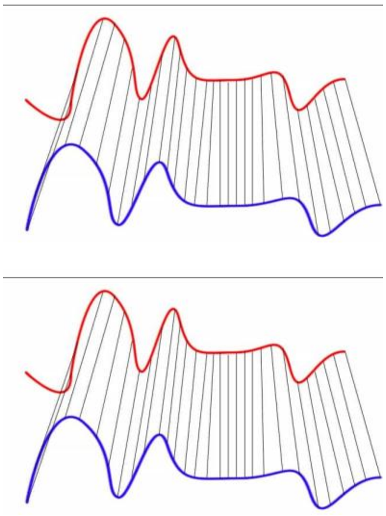

Fig 5: Speech Wave Alignment

The two images above represent the basic speech or sound wave alignment that is achieved using DTW. The DTW algorithm is used to align two vector sequences by turning the time axis repeatedly until the optimal match between the two sequences is found. The two sequences are aligned on the sides of the box, with one above and other on the left side. Both the sequences start at the bottom left of the grid.[8]

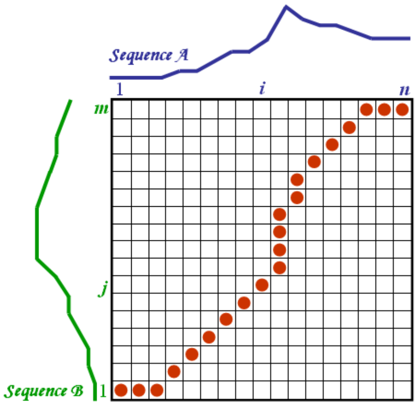

Fig 6:Scatter Diagram of Matching Sequence

In each cell, a measure of distance is plotted, which compares the corresponding elements of two sequences. The best match or alignment between these two sequences is the path through the grid, which minimizes the total distance between the two, which is called the global distance. DTW can be introduced based on two concepts.[8]

(i) Symmetrical DTW: Speech always depends on time as because it is. There are several ways of pronouncing the same word that can have different time time duration and also utterance of the same word with equal duration will deviate from it's middle. This is known as Dynamic Programming (DP). DP can find the minimum distance path through the matrix and can also help reduce the amount of calculation.

(ii) Asymmetrical DTW: In this approach, the input pattern which is in each frame is used only one time. It means that dispense and template length normalization and for diagonal transition no need to add local distance twice. This method is called asymmetric DP.[8]

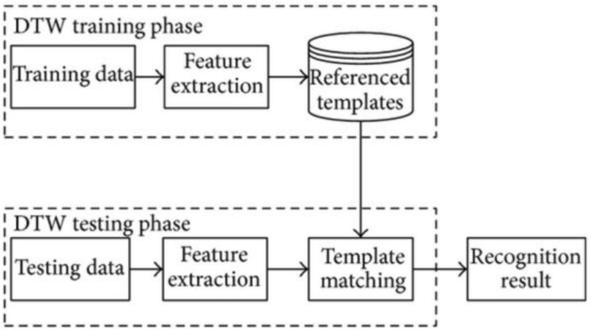

Fig 7:Block diagram for DTW

Advantages:

1. Increased recognition speed

2. Reduced storing space for reference template

3. Increased recognition rate

4. A threshold can be set which can be used to stop the process if the error is substantial.

Disadvantages:

1. To find the best reference template for a specific word, choosing the appropriate reference template is a difficult task.[13]

2.3 Hidden Markov Model (HMM): The Hidden Markov Model (HMM) is one of the most popular techniques in ML and statistics modeling sequences speech in the field of natural language and speech recognition. The speech which enters can be recognized mathematically using this approach. This is a doubly embedded stochastic (Having a random probability distribution or pattern that may be analyzed statistically but not be predicted precisely) process with cannot directly observable(hidden) stochastic process. HMM defines probability distribution for a set of observations a= a1,..., at,..., aT by talking another set of unobserved (hidden) discreet state variables u= u1,..., ut,..., uT. The main idea behind HMM is that the set of hidden states has Markov dynamics. Given ut, uT is independent of up for all T<t<p and that the observations at are independent of all other variables given ut. The model is defined by using two sets of parameters, the transition matrix whose ij element is P(ut+1=j | ut=i) and the emission matrix whose iq element is

P(at=q | ut=i). By using probabilistic model, stochastic modelling deals with incomplete and uncertain data. Incompleteness and uncertainty are common in speech recognition. For example, speaker variability, confusable sounds, contextual effects, etc. [5]

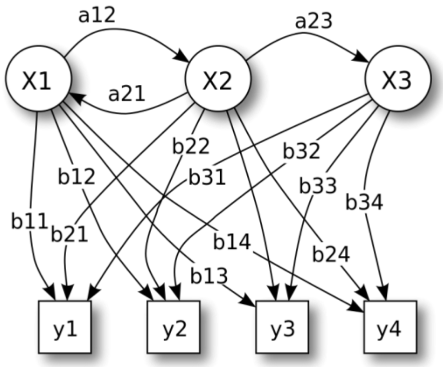

Fig 8:Working of HMM

Above Diagram represents the probabilistic model of a Hidden Markov Model, where X1, X2, X3 are the observable states and y1, y2, y3, y4 are the hidden states.

Advantages:

1. The HMM is very well structured algorithm which is probabilistic in nature hence, the algorithms are known for exact learning.

2. HMMs can represent the variance of the power demand of the appliance's using probability distribution.

3. HMMs can easily understand the dependencies and connections between two consecutive measurements.

Disadvantages:

1. Since the devices have a Markovian nature, they do not consider the state sequence that lead to into a given state.

2. Only the observed power feature are captured and not the others, apart from a few exceptions.

3. The dependency between appliances cannot be observed. However, conditional HMMs can capture some.[15]

2.4 N-grams: N-grams is the simplest type of language model. This model basically assigns probabilities to sentences or phrases. An N-gram model is basically a probability distribution based on the *n*th order Markov assumption. N- grams is a contiguous sequence of 'n' items from a given sample or text. The items can be anything from phonemes, syllables, letters, or words. The n-grams are usually collected from a speech or text corpus. The probability is calculated of the items.[2]

Let us understand this through an example. Let us take into consideration a series of sentences.

This is the *house* *that Jack built.*

This is the malt

That lay in the house that Jack built.

This is the rat

That ate the malt

That lay in the house that Jack built.

This is the cat*,*

That killed the rat,

That ate the malt,

That lay in the house that Jack built

The probability of any of the syllables, words or letters can be calculated with respect to any of the other syllables, words or letters.

Let's say p(house|this is the…)= p(this is the house)/p(this is the)= ¼

Therefore, the sentences containing the specific words "this is the house" go into the numerator and the sentences containing the words "this is the…" irrespective of what comes before or after it goes into the denominator. Therefore number of sentences containing "this is the…" are 4 and the number of sentences containing "this is the house" is 1 (As clearly highlighted in the examples above). Hence, the probability of this example is 1/4. These are termed as 'n-grams'. This algorithm can be used for single letters as well as complete sentences as per the requirement. N-grams is widely used in machine learning and deep learning field and plays a vital role in speech recognition too, speech recognition is a part of machine learning.

4. TOOLS TO OVERCOME CHALLENGES FACED WITH SPEECH RECOGNITION SYSTEMS

A number of noise reduction techniques have been engineered to extenuate the effect of noise on systems performance and often require the estimate of noise statistics. One technique that has been engineered is to design a database that may be used for the evaluation of feature extraction at the front end using a defined a Hidden Markov Model employed at backend.

4.1 Voice Activity Detector

Voice Activity Detector is a useful technique for enhancing the performance of Speech Recognition Systems employed in noisy environmental conditions [12]. Voice Activity Detector is used in Speed Recognition Systems for feature extraction process thereby resulting in enhancement of speech by the systems. The dependency factor of Voice Activity Detector lies on pitch detection, energy threshold, periodicity measure, and spectrum analysis. One major challenge that the detector face is while making a decision about extraction of Feature Vector (FV); selection of feature vector for signal detection and strong decision rule is a challenging problem, affecting the performance rate of Speech Recognition Systems.

4.2 AURORA Experimental Framework

The AURORA framework was designed as a contribution to the ETSI STQAURORA DSR Working Group. AURORA is developing standards for Distributed Speech Recognition (DSR) where speech analysis is done in telecommunication terminal and the recognition is performed at the central location in the telecom network [12]. In AURORA framework the idea was to design a database that may either be used for feature extraction along with backend defined Hidden Markov Model in Speech Recognition Systems. The TIDigits Database is used as the basis to develop the original database in AURORA framework.

5. Gap between machine and human speech recognition

What we know about human speech processing is still very limited, and we have yet to witness a complete and worthwhile unification of the science and technology of speech. In 1994, Moore [11] presented the following 20 themes which is believed to be an important to the greater understanding of the nature of speech and mechanisms of speech pattern processing in general:

a. How important is the communicative nature of speech?

b. Is human-human speech communication relevant to human machine communication by speech?

c. Speech technology or speech science? (How can we integrate speech science and technology).Whither a unified theory?

d. Is speech special?

e. Why is speech contrastive?

f. Is there random variability in speech?

g. How important is individuality?

h. Is disfluency normal?

i. How much effort does speech need?

j. What is a good architecture (for speech processes)?

k. What are suitable levels of representation?

l. What are the units?

m. What is the formalism?

n. How important are the physiological mechanisms?

o. Is time-frame based speech analysis sufficient?

p. How important is adaptation?

q. What are the mechanisms for learning?

r. What is speech good for?

How good is speech. After more than 10 years, we still do not have clear answers to these 20 questions.

6. CONCLUSIONS

The paper has focused on the basic speech recognition idea and it's working along with the algorithms like PLP, DWT and HMM, since, humans do a daily activity of speech recognition. Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. There are certain challenges faced by speech recognition devices and these challenges are overcome using Voice Activity detector and AURORA Experimental Framework. Speech recognition has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human machine interaction. We hope this paper brings about understanding and inspiration among the research communities of Speech Recognition.

ACKNOWLEDGEMENT

We wish to express our deepest gratitude to Prof.Suvarna Pansambal, Head of Department of Computer Engineering at Atharva College of Engineering for continuously motivating us to excel in the field of Computer Engineering and for guiding us by providing reviews and the necessary technical help regarding the information presented in this document.

REFERENCES

[1] IBM cloud education," Speech recognition ,"September 2020,

https://searchcustomerexperience.techtarget.com/definition/speech-recognition

[2]IBM cloud education, "Speech recognition, "September 2020,

https://www.ibm.com/cloud/learn/speech-recognition#toc-what-is-sp-xkYWoRbw

[3]Jonathan Hui, Speech Recognition-Feature extraction MFCC & PLP", August 2019

[4]https://machinelearning.conferenceseries.com/events-list/natural-language-processing- nlp-and-speech-recognition#:~:text=Natural%20Language%20Processing%20(NLP)%20is,with%20people%20using%20everyday%20language.

[5] Ravindu Senaratne, "The Promise of Natural Language Processing in Speech Recognition, "May 2017".

https://medium.com/ai-in-plain-english/the-promise-of-natural-language-processing-in-speech-recognition-dfdc4d276ba1

[6] Hermansky, Hynek, "Perpceptual linear predictive (PLP) analysis of speech, "The Journal of the Acoustical Society of America 87.4(1990): 1738-1752.

[7] Frederic Van Haren, "Natural Language Processing (NLP): meaningful advancements with BERT, "December 2019".

https://www.connect-converge.com/natural-language-processing-nlp-meaningful-advancements-with-bert/

[8] Permanasari, Yurika, Erwin H. Harahap and Erwin Prayoga Ali."Speech Recognition using Dynamic Time Warping (DTW)."Journal of Physics: Conference Series. Vol .1366, No. 1.IOP Publishing,2019.

[9] Reddy. .D. Raj. "Speech Recognition by machine: A review. "Proceedings of the IEEE 64.4(1976): 501-531.

[10] Yang Liu et.al,. Enriching Speech Recognition with Automatic Detection of sentence Boundaries an disfluencies , IEEE Transactions on Audio,Speech and Language processing, V.14,No.4,July 2006.

[11] Ramírez, J, Górriz, J.M., Segura, J.C. 2007. Voice Activity Detection, Fundamentals and Speech Recognition Systems Robustness, University of Granada, Spain.

[12] Hirsch, H.G., Pearce, D. 2000. The Aurora Experimental framework for the performance evaluation of Speech Recognition System under noisy conditions, ASR-2000 Automatic Speech Recognition: Challenges for the new Millennium Paris, France, 181-188.

[13]Smita B Magre, Ratnadeep R Deshmukh, "A comparative study of feature extraction techniques in speech recognition", May 2013, https://www.researchgate.net/profile/Ratnadeep_Deshmukh/publication/278549945_A_Comparative_Study_on_Feature_Extraction_Techniques_in_Speech_Recognition/links/5581520008aea3d7096e81c4/A-Comparative-Study-on-Feature-Extraction-Techniques-in-Speech-Recognition.pdf

[14]"NLP(Natural Language Processing) Tutorial, What is, History, Example", https://www.guru99.com/nlp-tutorial.html

[15]Oliver Parson, "The pros and cons of using HMMs to model appliances", May 2013, http://blog.oliverparson.co.uk/2013/05/the-pros-and-cons-of-using-hmms-to.html