Alternatives to Hidden Markov Model in Speech Recognition

Benjamin Allain, Colin Bontemps
May 17, 2011

Term Paper
DT2118 Speech and Speaker Recognition
KTH, School of Computer Science and Communication

Abstract

This paper has been written for the course DT2118 Speech and Speaker Recognition. It explores different alternatives and improvements to Hidden Markov Model (HMM) for automatic speech recognition (ASR). First we stress the weaknesses of the HMM methods, some of which are restricting. Then the focus is on hybrid systems where HMM methods are joined with other advanced techniques, particularly the Hidden Semi-Markov Model and Artificial Neural Networks some of which are promising. Afterwards we explain in details a completely alternative method called Landmarks based approach. It is still to be developed further but produces already really promising results. Finally we present briefly some others techniques for which we could not get enough documentation.

Introduction

During the past decades, important improvements have been made in the field of automatic speech recognition [8]. Hidden Markov Model turned out to be the most successful tool fur this purpose, and is used in most of the state-of-the-art systems.

These systems performances are still far from those of the human brain, and besides its popularity, the HMM has some weaknesses and limitations, which are detailed in the section I. In order to improve the performances of ASR systems and to overcome these limitations, several ways have been explored by the researchers. A first way is to improve the HMM system by using them simultaneously with other technologies, creating in this way new hybrid systems. Some of the most famous and/or promising are presented in section II. Some other researchers tried other ways to solves the problem of ASR. Section III presents in details a new paradigm, called Landmark-based approach, that competes with the HMM in terms of accuracy and computation speed. Section IV presents a few other solutions that we heard about but we haven't studied in details.

I. HMMs limitations

A. Duration Modeling

According to a HMM, the duration of a sound (*i.e.* how many times the process loops in one state) has a geometrical distribution. This distribution is not a good model of the actual duration of a phoneme/syllabe [6] (*cf* figure 1). The main problem is that the HMM overestimates the short durations probability. It is a problem for languages which distinguish short and long vowels.

B. First Order Assumption

The HMM is based on a first order Markov chain, which assumes that a state only depends on the previous state. It is possible to work with second order HMM, but the computational cost is huge, and the accuracy gain is small. So this limitation is not a crucial problem.

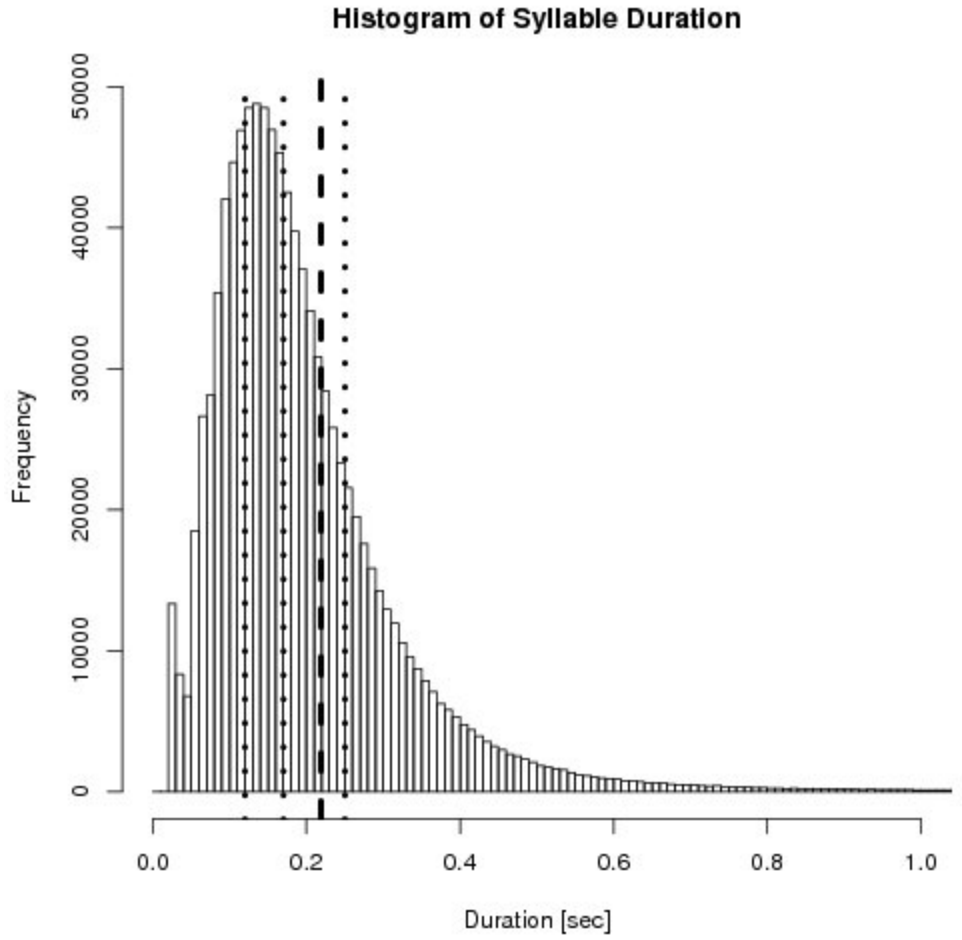

Figure 1 : Histogram of syllabe. Duration. Below 0.1 second, the exponential model is not viable. From [9]

C. Conditionnal Independence Assumption

The HMM assumes that the observations only depends on the underlying states. This is wrong, because there are also correlated to each other. For example, when a phoneme occurs twice in an utterance, the two occurrences are very likely to be similar, because the speaker is the same, in the same mood, and the capture context (room, microphone, ambient noise) are the same. For the same reasons, and because one sound often covers more than one frame, consecutive frames are also likely to be similar. Thus the HMM loses some useful information here.

II. Hybrid Systems

To overcome the limitations of HMM modeling, but still taking advantages of it's low computational cost, the HMM has been modified to create hybrid systems.

A. Hidden Semi-Markov Model

This approach [6] consists in keeping the Markov chain transition model between different states, but the transition model between a state and itself is replaced by a more appropriate model, in order to fit better the actual duration distribution. The duration model is still parametric (often a gamma distribution). The process is no longer a Markov Chain (the first order assumption is violated), so the Viterbi algorithm needs to be modified. Therefore, the time consumption is D times higher (where D is the maximum number of state repetition).

The accuracy in recognition is improved by 8% in term of LER (Letter Error Rate) in a large vocabulary continuous speech recognition task (not detailed in the reference) in Finnish (this language distinguishes certain phones only with the duration).

B. HMM/ANN Hybrid Systems

It is possible to use a Artificial Neural Network (ANN) for phoneme classification given a segment of consecutive frames. The solution proposed here [5] uses a Multi-layer Perceptron (MLP) with a fixed number of inputs (frames) and one outputs for each phoneme (likelihood). To adapt to any segment length, the segment is sub-sampled (nearest neighbour) if necessary. To take into account the segment length, the segment length is also an input of the network. Unlike HMMs, the MLP is able to take advantage of the frames interdepedencies.

Given a segmented utterance and a corresponding phonetic transcription (hypothese), the likelihood of each segment are computed by the MLP, then multiplied to form the likelihood of the transcription given the utterance.

However, the MLP needs a segmentation. The number of possible segmentation is too enormous to generate all of them. That's why a HMM is used to generate the N best segmentations. Then the MLP is used to score each of these segmentations. The N segmentations are then resorted according to a linear combination of the HMM score and the MLP score.

This ANN/HMM hybrid system decreases by 20% the error rate with N=20 compared to the HMM alone [5]. The test data consisted of 3990 utterances from 109 speakers (from the DARPA 1000-words Resource Management speech corpus).

III. Landmark-based Approach

A. Introduction

The Landmark-based speech recognition gives an approach to the problem of speech recognition which includes more information about the speech production process. This approach tends to first segment and classify the audio within different types of sound (vowels, nasal, fricative) and then processes an appropriate analysis for each segment. In some cases it has given results comparable to these from state-of-the-art HMM machine.

B. Process

B.1. Segmentation of the Audio

a. Categorization

The best way to explain what is considered as a feature in the case of this approach is to give examples. Here a so called feature can be “sonorant” or “nasal”. This feature are given binaries value so that a sound can be “+sonorant” and “+nasal”. There are organised in a hierarchy, and can be represented on a tree. Here is an example of such a classification tree in Figure 2 from [4] :

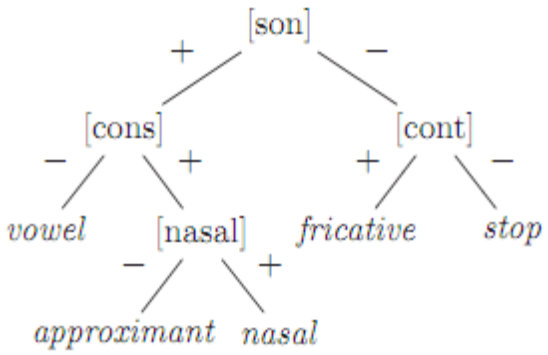

Figure 2: The hierarchy of distinctive features leading to the classes. From [4]

This tree can be more complex as the one shown in [3]. The point is that every phone can be classified in such a tree. This different features are the one used to proceed the segmentation.

b. Features Extraction

The point of the feature extraction is to map every feature to a continuous parameter. As every feature is different the mapping to this parameter will also be different. This mapping has to be coherent with the binary information such as if the mapped parameter is positive a time t, the corresponding feature will have for binary value “feature+”. Here a strong link with the articulation may be found. For example if one gets the parameter “sonorant” from the repartition of the energy within the spectrum and the local periodicity of the signal, it will be more or less mapped to the opening of the vocal tract. This analogies can be brought further, with most of the features, and especially as the tree go more complex. For example, if one subdivides the feature vowel in “front” and “back”, it can then map their associated parameter to the positioning of the tongue and the mouth. The point is that for extracting each features parameter, different aspects of the signal will be considered. The information extracted will then take the form of different observation functions which give the value of each parameter at each time in the sampler.

c. Segmentation Process

The aim of the segmentation is to identify the changes in the articulation in order to segment the sampler. For each segment is associated the binary feature so that it is classified in an unique class such, i.e. : a leaf of the classification tree.
In order to accomplish the segmentation, the only information that will be used is extracted from the observation function, such as their minima, maxima, inflexion points…

The first segmentation will be done using the feature “sonorant”. The sampler is divided in sonorant and non sonorant parts. Then in every sonorant part, the different vowels are counted. If a sonorant segment has N vowels, N+1 intervocalic segment will be inserted.

Once this first cut is done, the different segment can be deeper analysed to be classified. In the probabilistic analysis, they can be considered as independent one from each other (as in [4]) or be proceed with an higher order analysis.

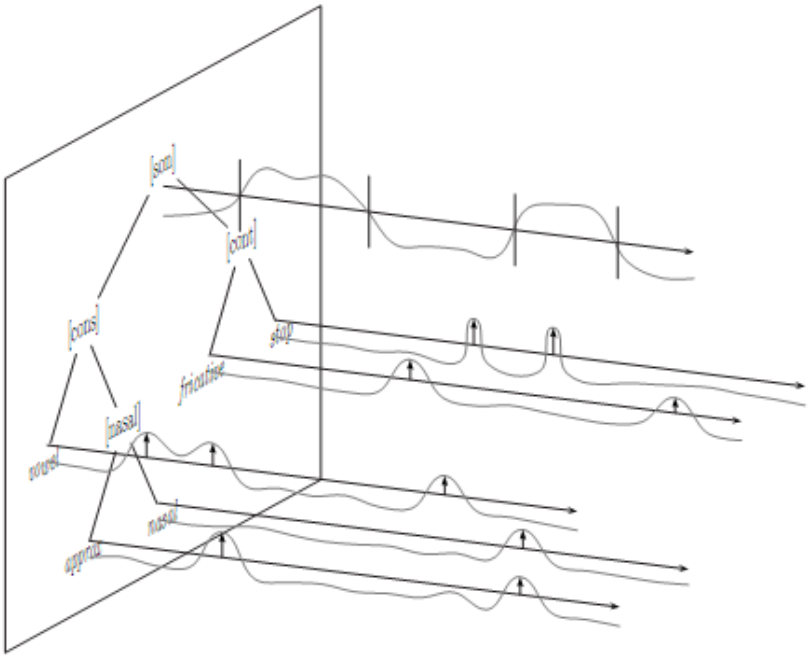

Figure 3 : Schematic diagram of parameter representation in the tree. Landmarks are indicated by vertical arrows. From [4]

B.2. Recognition with Custom Parameters and Methods

If the segmentation gave good results, the recognition of the sounds is then much easier. Knowing the length and the class of a segment, it can be analyzed with appropriate methods and parameters. For examples if a 100ms long segment has been analysed as a vowel, a formant analysis can be done, or if a 50ms long segment is classified as a fricative, it can the be further analysed in order to know which fricative exactly.The segmentation previously made allow to adapt the time windows to fit the segments, which avoid a lot of computing when it is not needed.

C. Comparison with HMM Methods

C.1. Results Comparison

Train/Test	Ad/Ad	Fe/Ma	Ma/Fe	Ch/Ch	Ad/Ch	Ch/Ad
39 MFCCs	99.88	68.29	70.27	98.30	60.20	62.37
30 APs	99.53	79.24	90.90	97.50	85.70	89.81

Table 1: Digit recognition accuracy (%) (TI46 corpus for Adults, TIDIGITS corpus for children). Ad = Adult, Ch = Child, Fe = Female, Ma = Male. From [3]

As shown in *Table 1,* the Landmark-based approach gives results comparable to these of HMM on a simple task such as digit recognition.

C.2. Advantages on HMM

First, as we can see in the table 1 this method gives good results with speaker independent recognition involving different kind of person. Then it has also been proved that the Landmark method is much easier to adapt to another language (indeed the basic classification does not depend a lot of the language), and gives faster better results. Indeed the structure of the different sounds of the speech is similar in a lot of countries, even if further in the classification they differ. For example the vowels or the plosives are not exactly the same, but there is still this vowels and plosives groups. Finally the categorisation previously done allows to spare computation time when analyzing the segments.

C.3. limitations

Being the strong point of this method, the segmentation is also its weak point. If this segmentation is not done properly, we cannot expect a good recognition at the end. Because of this the Landmark method is also more vulnerable to noise than HMM.

IV. Other methods

For the following alternatives to HMM we have not found enough references or explications to treat them in details, but we think there are worth to be mentioned this paper.

A. Hidden Bernoulli Model

In [2] an alternative to HMM is presented: Time-Inhomogeneous Hidden Bernoulli Model (TI-HBM). It is said to have comparable results to HMM with simpler methods and less computing for both training and recognition, especially when the number of Gaussian Mixtures is low. In TI-HBM, the state transition process is a generalized Bernoulli process instead of a Markov one.

B. Linear Dynamic Model

The principle of this model [4] is to use the HMM with a continuous-valuated hidden state instead of the usual discrete-valuated hidden state. According to the authors, it relaxes the conditional independence assumption. The performances are better than regular HMM for monophones, but not for triphones. It is also

computationally expansive.

C. General Regression Neural Networks

An article [7] proposes to use a neural network performing a general regression directly on the features vectors. This technique doesn't use the HMM at all. It turns out to give a better WER than HMM and HMM/MLP hybrid systems.

Conclusion

To overcome the HMM limitations, the alternatives of HMM can be divided in two categories. The ones who modify the HMM: HSMM, ANN/HMM hybrid and Linear Dynamic Model; and the ones who are intrinsically different: Landmark-based approach, Hidden Bernouilli Model and General Regression Neural Networks. Some of them are quite promising, even if not all of them are completely developed yet. These different techniques will hopefully allow to improve further the ASR systems and make them ready for more practical uses.

References

[1] Leutnant, Volker / Haeb-Umbach, Reinhold (2010): On the exploitation of hidden Markov models and linear dynamic models in a hybrid decoder architecture for continuous speech recognition, In *INTERSPEECH-2010*, 2946-2949.

[2] Jahanshah Kabudian, M. Mehdi Homayounpour, S. Mohammad Ahadi : Time-Inhomogeneous Hidden Bernoulli Model : an Alternative to Hidden Markov Model for Automatic Speech Recognition, 2008, Department of Computer Engineering, Department of Electrical Engineering, AmirKabir University of Technology, Tehran.

[3] Carol Y. Espy-Wilson, Tarun Pruthi, Amit Juneja, Om Deshmukh : Landmark-based Approach to Speech Recognition: An Alternative to HMMs, 2007, Institute for Systems Research and Dept. of Electrical and Computer Eng., University of Maryland

[4] Aren Jansen and Partha Niyogi : Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition, 2007, Department of Computer Science, University of Chicago

[5] S. Austin, G. Zavaliagkos t , J. Makhoul, and R. Schwartz, Improving State-of-the-Art Continuous Speech Recognition Systems Using the N-Best Paradigm with Neural Networks, 1992, BBN Systems and Technologies, Cambridge, Northeastern University, Boston

[6] Janne Pylkkönnen and Mikko Kurimo, Duration Modeling Techniques for Continuous Speech Recognition, Neural Networks Research Centre Helsinki University of Technology, Finland, 2003

[7] Abderrahmane Amrouche, and Jean Michel Rouvaen, Efficient System for Speech Recognition using General Regression Neural Network, 2006, International Journal of Intelligent Technology Volume 1 Number 2

[8] The History of Automatic Speech Recognition Evaluations at NIST, http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html, 2009

[9] Florian Schiel, Statistics of Conversational German, 2010, Ludwig maximilians universität München