Isolated Digit Speech Recognition Based On Hidden Markov Models

Faculté des Sciences et Techniques - Mohammedia
جامعة الحسن الثاني بالدار البيضاء
Université Hassan II de Casablanca
كلية العلوم و التقنيات - المحمدية

Fatima Zahra EL FATEHY, Mohammed KHALIL, and Abdellah ADIB

Team Networks, Telecoms
Multimedia, LIM@II-FSTM, B.P. 146, 20650
Mohammedia, Morocco
fzelfatehy@gmail.com, medkhalil87@gmail.com, adib@fstm.ac.ma

Abstract

In the last decades, the Automatic Speech Recognition (ASR) technologies have improved and nowadays are present in various domains of application. In this work, we present an isolated digit Speech Recognition System, which recognizes a spoken digits in English. The speech signal related to a word is first submitted to a feature extraction step using Mel Frequency Cepstral Coefficient (MFCC). Hidden Markov Models (HMM) are then used to model the utterances, with an observation likelihood modeled by a Gaussian Mixture Model (GMM). Finally, Viterbi search algorithm is used for the decoding process.

Keywords: Automatic Speech Recognition, isolated digit, Hidden Markov Model, Gaussian Mixture Model

Introduction

Digit speech recognition systems are used in different applications, such as phone dialing, banking systems, catalog-dialing etc. In English, They have been developed using the Weighted Mel Frequency Cepstral Coefficients (WMFCC) and an Improved Features for Dynamic Time Warping (IFDTW) algorithm [2], and using Hidden Markov Model (HMM) [10]. They have been investigated in Arabic, by applying the Harmonic Noise Model (HNM) algorithm to extract the acoustic parameters, which are trained and recognized by the HMM [5]. In Japanese using Learning Vector Quantization (LVQ) [7]. And Kannada [9] where Mel Frequency Cepstral Coefficients (MFCC) are used as the features and HMM as feature recognizer, and to obtain the observation sequence K-means procedure is performed on the feature vectors.

In this work, we focus on isolated digit recognition based on HMMs and Gaussian Mixture Model (GMM). We evaluated the system in order to find the best feature vectors to represent the samples and the number of GMM classes used to estimate the observation probabilities.

Conclusion

In this paper, we worked on an Isolated Digit Speech Recognition System using HMMs and GMMs. Experimental results showed the best choice of the number of classes in a GMM per state when using different feature vectors.

The system description:

The system is composed of two phases. A training phase where the model parameters are learned from labeled speech signals and a decoding phase that finds the most likely spoken word given speech features using the learned models. The first step in both phases consists of extracting the linguistic information in the speech signal.

First, the signal is windowed using a Hamming window [6] to reduce discontinuities. The MFCC [4] is computed afterwards for each frame by first applying Fourier transform. In order to adapt the frequency resolution to the properties of the human ear, the power spectrum is next integrated within an overlapping windows called Mel filters. Then, a discrete cosine transformation is applied to the logarithm of the Mel filters outputs. In order to improve the performance of ASR systems based on HMMs, first (ΔMFCC) and second (ΔΔMFCC) time derivatives are added to the features vector [8].

To model the speech signal into small units, we use HMM. An HMM corresponds to an N-state probabilistic automaton with a hidden process defined by states and transitions, which represents the words that are the underlying source of the acoustics in the observed process [11]. An HMM is specified by a set of states Q along with a set of transition probabilities A with elements a_{ij}, a defined start state and end state, a set of observation O, a probability distribution for each state $B = b_i(o_t)$ and a prior probability distribution over states $\pi = \{\pi_1, \pi_2, \ldots, \pi_\mathrm{N}\}$. The observation likelihoods are estimated in continuous distribution by GMMs [3] which is defined as:

$$b_i(o_t) = \sum_{k=1}^{K} w_k\, \mathcal{N}(x{:}\, \mu_k, \Sigma_k)$$

where w_k is a mixture weight, with $\sum_{k=1}^{K} w_k$=1 and $\mathcal{N}(x{:}\, \mu_k, \Sigma_k)$ is a Gaussian distribution. Each state of the HMM has its own GMM. Figure 1 shows an HMM-GMM model.

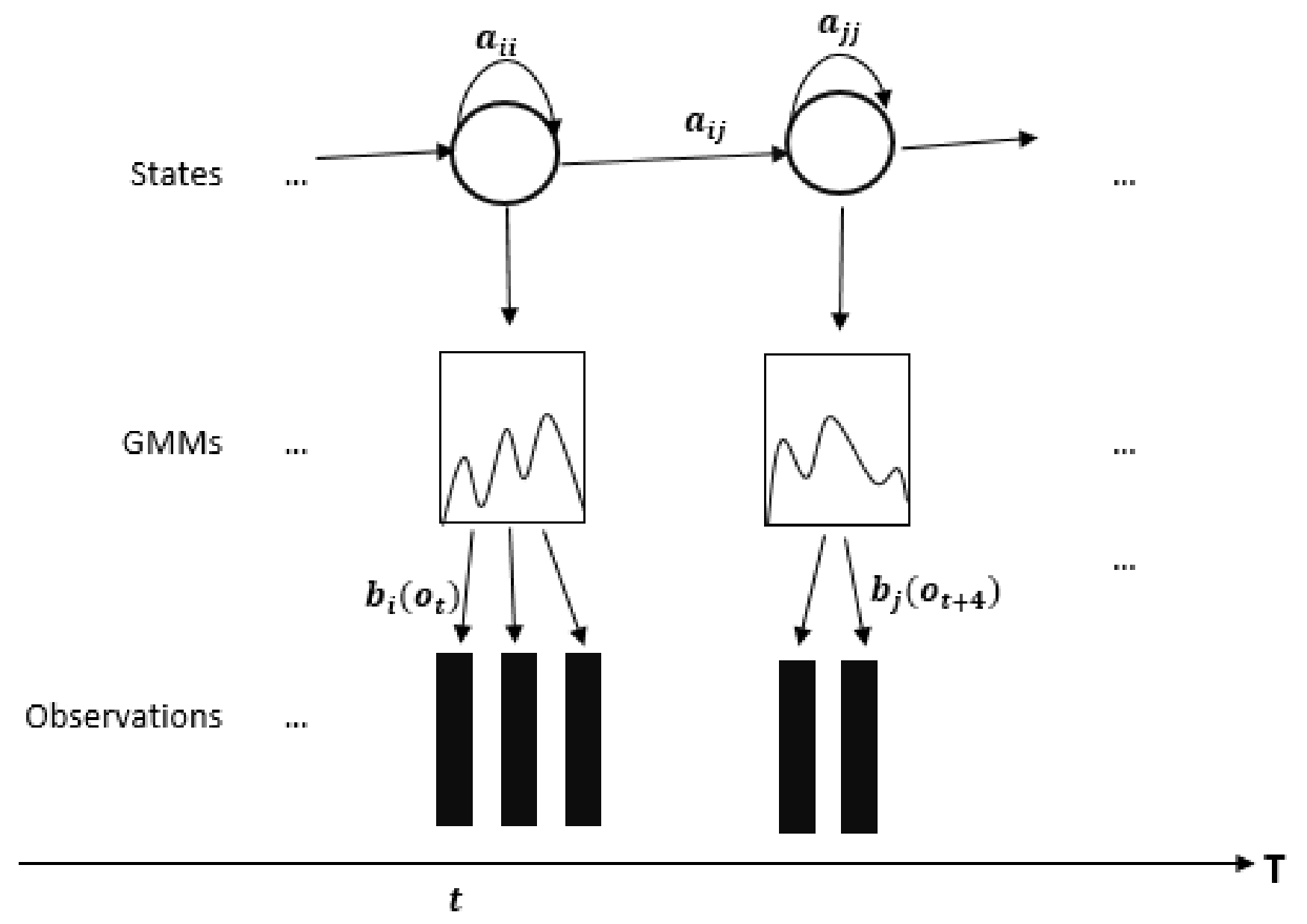

Figure 1: an HMM-GMM model.

In the training phase, the HMM models are trained using a dataset. A transcription is associated with each sound sample so as an HMM can be modeled by its representatives in the corpus. Thus, the goal of training is to estimate the optimal parameters of the HMM of each unit, it is based on the Maximum Likelihood Estimation (MLE) criterion estimated by the Baum-Welch algorithm [1]. The estimation consists in determining all the transition probabilities A as well as the means μ, variances $\sum$ and mixture weights $w$of all GMMs.

For the decoding phase, we need to deduce the sequence of states that generated the given observations. This task is accomplished through the Viterbi search algorithm [12] using the probabilities generated by the HMM models. The principle is to iteratively build the most likely sequence of states by tracing back to the beginning the best path which maximizes the likelihood at state i at time T.

Results

In this section, we evaluate the results of the system. In this work, we used ten distinct English digits (zero through nine). Each one of them in the speech dataset* is represented in the training by 45 samples and in the testing by 5 samples of the same speaker. The speech signal is sampled at 8 KHz, we used a hamming window of 25 ms with a step of 10 ms between two successive frames. The feature vectors dimensions (d) are set to 13, 26 and 39 using MFCC, ΔMFCC and ΔΔMFCC. The left-right HMMs are used to model the words, in which each state transition only to itself and the next state. The states of the HMM models represent the building blocks of the word known as phones, in addition to a start and end states. The observation likelihoods are estimated using GMMs, with diagonal matrices. The number of Gaussians (G) per state is set to 1, 2, 4, 8, 16 and 32.

In order to evaluate our isolated digit recognition system, we fixed the value of d while varying G. From the results listed in Figure 2. For d=13, the system achieves a 98% accuracy when using a single Gaussian for the observation likelihoods. While in d=26, the highest accuracy of 94% is reached for a 4 mixtures per state. In d=39, the use of the 8, 16 or 32 Gaussian by state reached an 84% accuracy.

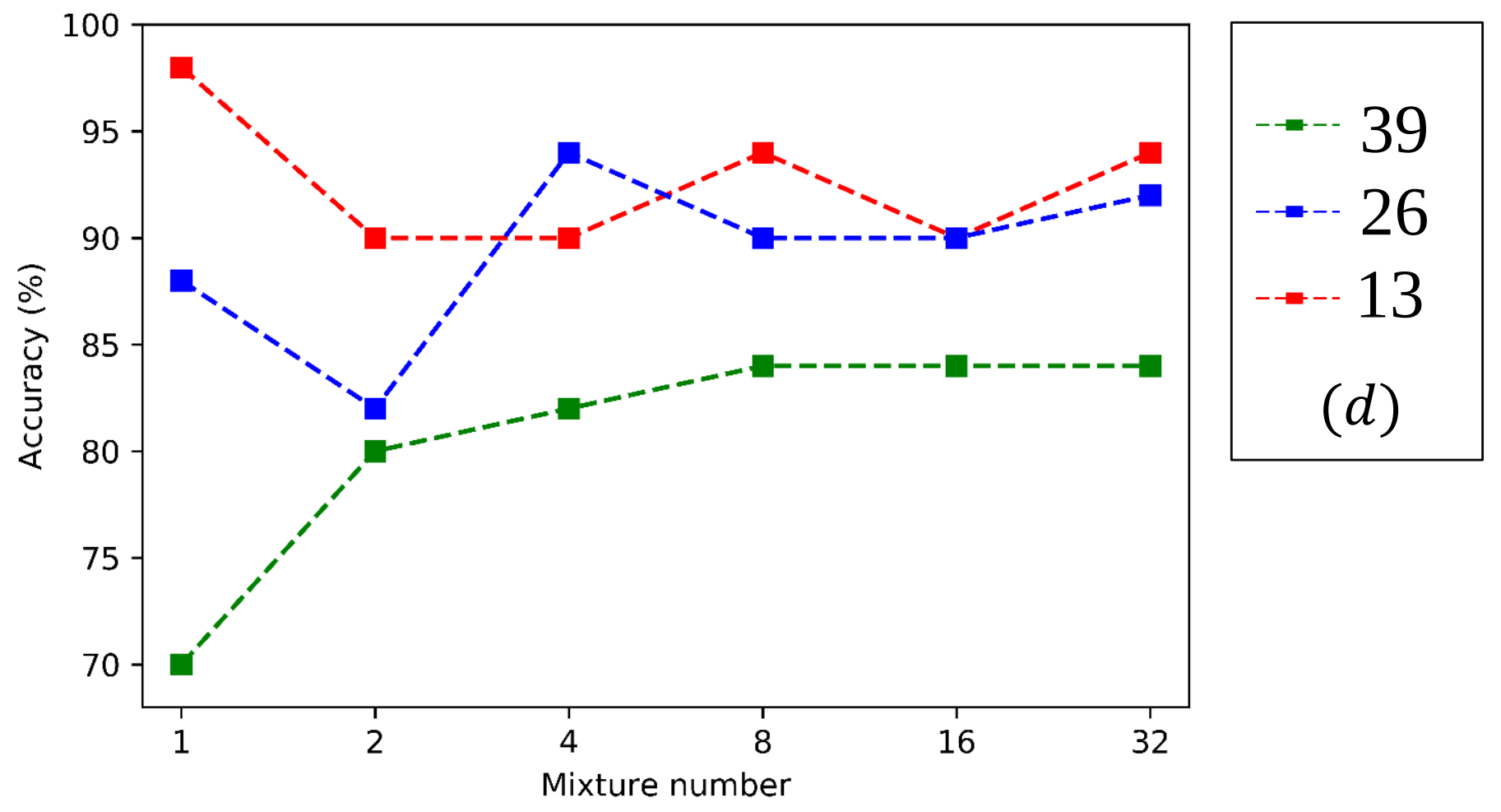

Figure 1: Recognition rates of the tested data.

* https://github.com/thayermldac/spoken-digit-dataset

References

[1] Leonard Baum. “An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process”. In: Inequalities 3 (1972), pp. 1–8.
[2] Santosh V Chapaneri and Deepak J Jayaswal. “Efficient speech recognition system for isolated digits”. In: IJCSET 4.3 (2013), pp. 228–236.
[3] George E Dahl et al. “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition”. In: IEEE Transactions on audio, speech, and language processing 20.1 (2012), pp. 30–42.
[4] Steven Davis and Paul Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: IEEE transactions on acoustics, speech, and signal processing 28.4 (1980), pp. 357–366.
[5] A Guerid, H Saboune, and A Houacine. “Recognition of isolated digits using HMM and harmonic noise model”. In: 2018 1st International Conference on Computer Applications & Information Security (ICCAIS). IEEE. 2018, pp. 1–5.
[6] Fredric J Harris. “On the use of windows for harmonic analysis with the discrete Fourier transform”. In: Proceedings of the IEEE 66.1 (1978), pp. 51–83.
[7] K Kondo, H Kamata, and Y Ishida. “Speaker-independent spoken digits recognition using LVQ”. In: Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94). Vol. 7. IEEE. 1994, pp. 4448–4451.
[8] K-F Lee and H-W Hon. “Speaker-independent phone recognition using hidden Markov models”. In: IEEE Transactions on Acoustics, Speech, and Signal Processing 37.11 (1989), pp. 1641–1648.
[9] H Muralikrishna, T Ananthakrishna, et al. “HMM based isolated Kannada digit recognition system using MFCC”. In: 2013 international conference on advances in computing, communications and informatics (ICACCI). IEEE. 2013, pp. 730–733.
[10] Ganesh S Pawar and Sunil S Morade. “Realization of Hidden Markov Model for English Digit Recognition”. In: International Journal of Computer Applications 98.17 (2014).
[11] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: Proceedings of the IEEE 77.2 (1989), pp. 257–286.
[12] Andrew Viterbi. “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm”. In: IEEE transactions on Information Theory 13.2 (1967), pp. 260–269.