Digit recognition using Continuous Density Hidden Markov Models for different speakers

José Luis Oropeza Rodríguez[1] and Sergio Suárez Guerra[1]

[1] Center for Computing Research, National Polytechnic Institute,
Juan de Dios Batiz esq Miguel Othon de Mendizabal s/n, P.O. 07038, Mexico
joropeza@cic.ipn.mx, ssuarez@cic.ipn.mx

Abstract. The goal of automatic speech recognition (ASR) is to develop techniques and systems that enable computer to accept speech input. The digit task has been so much employed to contribute the effort in ASR. Since 1950's, Automatic Speech Recognition Systems (ASRS) has been studied by so much researchers. In our country, universities have inverted a lot of time creating speech recognition systems (UNAM, IPN, ITESM, UAM, UDLA, among others). In this paper we present the results obtained when we trained a digit corpus with 5 people (3 men and 2 women). Firstly, we trained our system with 5 speakers, and we obtained a 94.13% of recognition rate. Secondly, we probed the ASR individually for each speaker and we obtained a recognition rate above of 98.5%. Thirdly, we employed a digit corpus with men only and another with women only, the results obtained were above of 95% for the men corpus and 98% for women corpus. Finally, we report that we used 300 sentences of speech signal (100 for training task and 200 for recognition task) from each speaker, then we processed for the complete corpus (5 speakers), a total of 1500 sentences were utilized. The results were obtained using Hidden Markov Model Toolkit (HTK), and personal software.

Keywords. Automatic Speech Recognition, speaker recognition, Continuous Density Hidden Markov Models, Viterbi Trainning, and dependent-independent speaker automatic speech recognition systems.

1 Introduction

Speech and language are perhaps the most evident expression of human thought and intelligence –the creation of machines that fully emulate this ability poses challenge that reach far beyond the present state of the art.

© A. Argüelles, J. L. Oropeza, O. Camacho, O. Espinosa (Eds.)
Computer Engineering.
Research in Computing Science 30, 2007, pp. 1-12

The speech recognition field has been fruitfully and productively benefited from sciences as diverse as computer science, electrical engineering, biology, psychology, linguistics, statistics, philosophy, physics and mathematics among others. The interplay between different intellectual concerns, scientific approaches, and models, and its potential impact in society make speech recognition one of the most challenging, stimulating, and exciting fields today.

As early as 1950s, simple recognizers have been built, yielding credible performance. But it was soon found that the techniques used in these systems were not easily extensible to more sophisticated systems. In particular, several dimensions emerged that introduce serious design difficulties or significantly degrade recognition performance.

Most notably, these dimensions include:

- Isolated, connected, and continuous speech
- Vocabulary size
- Task and language constraints
- Speaker dependence or independence
- Acoustic ambiguity, confusability
- Environmental noise

The first question one should ask about a recognizer or a task is: is the speech connected or spoken one word at a time? Continuous Speech Recognition (CSR) is considerably more difficult than isolated word recognition (IWR) that is because at first, word boundaries are typically not detectable in continuous speech, at second, there is much greater variability in continuous speech due to stronger coarticulation (or interphoneme effects) and poorer articulation ("El ave es grande" becomes "la vez").

A second dimension is the size of the vocabulary. Exhaustive search in very large vocabularies is typically unmanageable. Instead, one must turn to smaller sub-word units (phonemes, syllables, triphonemes, etc.), which may be more ambiguous and harder to detect and recognize.

A system with a semantic component may eliminate such sentences from consideration. A system with a probabilistic language model can effectively use this knowledge to rank sentences.

These knowledge sources or language models can reduce an impossible task to a trivial one. The challenge in language modeling is to derive a language model that provides maximum constraint while allowing maximum freedom of input. The constraining power of a model language can be measured by perplexity, roughly the average number of words that can occur at any decision point.

The different sources of variability that can affect speech determine most of difficulties of speech recognition. During speech production the movements of different articulators overlap in time for consecutive phonetic segments and interact with each other. As a consequence, the vocal tract configuration at any time is influenced by more than one phonetic segment. This phenomenon is known as coarticulation mentioned above. The principal effect of the coarticulation is that the same phoneme can have very different acoustic characteristics depending on the context in which it is uttered [Farnetani 97].

The most prominent issue is that of Speaker dependence as opposed to speaker independence. A speaker dependent system uses speech from the target speaker to learn its model parameters. On the other hand, a speaker-independent system is trained once and for all, and must model a variety of speaker's voice.

Speech recognition-system performance is also significantly affected by the acoustic confusability or ambiguity of the vocabulary to be recognized. A confusable vocabulary requires detailed high performance acoustic pattern analysis. Another source of recognition-system performance degradation can be described as variability and noise.

Finally, the applications of the ASR are vast, for example: Credit-card numbers, telephone numbers, and zip codes, require only a small vocabulary.

2 Characteristics and Generalities

"The schools of thought in speech recognition" describe four different approach researched at today, they are [Kirschning 1998]:

- template-based approach
- knowledge-based approach
- stochastic approach and,
- connectionist approach

Before continuing described the characteristics of them, we must to say that ASR has implemented one stage called "speech analysis". The applications that need voice processing (such as coding, synthesis, and recognition) require specific representations of speech information. For instance, the main requirement for speech recognition is the extraction of voice features, which may distinguish different phonemes of a language.

To decrease vocal message ambiguity, speech is therefore filtered before is arrives at the automatic recognizer. Hence, the filtering procedure can be considered as the first stage of speech analysis. Filtering is performed on discrete time quantized speech signals. Hence, the first procedure consists of a conversion analog to digital signal. Then, the extraction procedure of the significance features of speech signal is performed.

In the template-based approach, the units of speech (usually words, like in this work), are represented by templates in the same form as the speech input itself. Distance metrics are used to compare templates to find the best match, and dynamic programming is used to resolve the problem of temporal variability. Template-based approaches have been successful, particularly for simple applications requiring minimal overhead.

In the knowledge-based approach, proposed in the 1970s and early 1980s. The pure knowledge-based approach emulates human speech knowledge using expert systems. Rule-based systems have had only limited success. The addition of knowledge was found to improve other approaches substantially. Recently, in the Spanish language a new approach using the rules of the syllabic units has showed the utility of these units in the ASR.

The stochastic approach, which is similar to the template-based approach has been using in the recent developments of ASR. One major difference is that the probabilistic models (typically Hidden Markov Models –HMM-) are used. HMM are based on a sound probabilistic framework, which can model the uncertainty inherent in speech recognition. HMM have an integral framework for simultaneously solving the segmentation and the classification problem, which makes them particularly suitable for continuous-speech recognition. One characteristic of HMM is that they make certain assumptions about the structure of speech recognition, and then estimate system parameters as though the structures were correct.

The connectionist approach use distributed representations of many simple nodes, whose connections are trained to recognize speech. Connectionist approach is a most recent development in speech recognition. While no fully integrated large-scale connectionist systems have been demonstrated yet, recent research efforts have shown considerable promise. Some of the problems that remain to be overcome include reducing time training and better modelling of sequential constraints.

3 Automatic Speech Recognition Systems

The frequency bandwidth of a speech signal is about 16 KHz. However, most of speech energy is under 7 KHz. Speech bandwidth is generally reduced in recording. A speech signal is called orthophonic if all the spectral components over 16 KHz are discarded. A telephonic lower quality signal is obtained when ever a signal does not have energy out of the band 300-3400 Hz. Therefore, digital speech processing is usually performed by a frequency sampling ranging between 8000 samples/sec and 32000 samples/sec. These values correspond to a bandwidth of 4 kHz and 16 kHz respectively. In this work, we use a frequency sampling 11025 samples/sec [Bechetti and Prina 1999].

The excitation signal is assume periodic with a period equal to the pitch for vowels and other voice sounds, while for unvoiced consonants, the excitation is assumed white noise, i.e. a random signal without dominant frequencies. The excitation signal is subject to spectral modifications while it passes through the vocal tract that has an acoustic effect equivalent to linear time invariant filtering. These modifications give to the final sound the characteristic features of the different phonemes of a language. The model is relevant, for each type of excitation; a phoneme is identified mainly by considering the shape of the vocal tract configuration can be estimated by identifying the filtering performed by the vocal tract on the excitation. Introducing the power spectrum of the signal $P_x(\omega)$, of the excitation $P_y(\omega)$ and the spectrum of the vocal tract filter $P_h(\omega)$, we have:

$$P_x(\omega) = P_y(\omega)P_h(\omega) \qquad [1]$$

Where ω is the frequency of the discrete time signal. The spectrum of the filter can be obtained from the power spectrum of the speech $P_x(\omega)$ the contribution of the excitation power $P_y(\omega)$.

3.1 Signal preprocessing

The characteristics of the vocal tract define the current uttered phoneme. Such characteristics are evidenced in the frequency domain by the location of the formants, i.e. the peaks given the resonances of the vocal tract. Although possessing relevant information, high frequency formants have smaller amplitude with respect to low frequency formants. A preemphasis of high frequencies is therefore required to obtain similar amplitude for all formants. Such processing is usually obtained by filtering the speech signal with a first order FIR filter whose transfer function in the z-domain is [Oppenheim 89]:

$$H(z) = 1 - az^{-1} \qquad [2]$$

a being the preemphasis parameter. In essence, in the time domain, the preemphasized signal is related to the input signal by the relation:

$$x'(n) = x(n) - ax(n-1) \qquad [3]$$

A typical value for ***a*** is 0.95, which gives rise to a more than 20 dB amplification of the high frequency spectrum.

3.2 Windowing

Traditional methods for spectral evaluation are reliable in the case of a stationary signal (i.e. a signal whose statistical characteristics are invariant with respect to time). For voice, this holds only within the short time intervals of articulatory stability, during which a short time analysis can be performed by "windowing" a signal $x'(n)$ into a sequence of windowed sequences $x_t(n), t = 1,2,\ldots.,T$ called frames, which are then individually processed:

$$x_t^{'}(n) \equiv x'(n-t \cdot Q), \quad 0 \leq n < N, \quad 1 \leq t \leq T \qquad [4]$$

$$x_t(n) \equiv w(n) \cdot x_t^{'}(n) \qquad [5]$$

Where $w(n)$ is the impulse response of the window. Each frame is shifted by a temporal length Q. If $Q=N$, frames do not temporally overlap while if $Q<N$, N-Q samples at the end of a frame $x_t^{'}(n)$ are duplicated at the beginning of the following frame .

In ASR, the most-used window shape is the Hamming window, whose impulse response is a raised cosine impulse [DeFatta et al. 1988]:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) & n = 0,..,N-1 \\ 0 & otherwise \end{cases} \qquad [6]$$

4 Hidden Markov Models and Experimental Methodology

Now, we are going to show the algorithms employed for Automatic Speech Recognition using Hidden Markov Models (HMMs). Like we know, HMMs mathematical tool applied for speech recognition presents three basic problems [Rabiner and Biing-Hwang, 1993] y [Zhang 1999]:

Problem 1. Given the observation sequence $O = O_1 O_2 \ldots O_T$, and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?

1. Initialization

$$\alpha_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N \qquad [7]$$

2. Induction

$$\alpha_{t+1}(j) = b_j(O_{t+1}) \sum_{i=1}^{N} \alpha_t(i) a_{ij} \quad 1<j\leq N,\ 1\leq t\leq T-1 \qquad [8]$$

3. Termination

$$P(O \mid \lambda) = \sum^{N} \alpha_T(i) \qquad [9]$$

Problem 2. Given the observation sequence $O = O_1 O_2 .... O_T$ and the model λ, how do we choose a corresponding state sequence $Q = q_1 q_2 ... q_T$ which is optimal in some meaningful sense?

1. Initialization

$$\delta_1(i) = \pi_i b_i(O_1) \quad 1 \le i \le N \qquad [10]$$

2. Recursion

$$\delta_{t+1}(j) = b_j(O_{t+1})\left[\max_{1\le i\le N} \delta_t(i) a_{ij}\right] 1\le j\le N, 1\le t\le T\text{-}1 \qquad [11]$$

3. Termination

$$p^* = \max[\delta_T(i)] \quad 1 \le i \le N \qquad [12]$$

$$q^* = \arg\max[\delta_T(i)] \quad 1 \le i \le N \qquad [13]$$

Problem 3. How do we adjust the model parameters
$\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$?

$$a_{ij} = \frac{\exp\, ected \quad number \quad of \; times \; from \; state \; s_i to \; state \; s_j}{\exp\, ected \; number \; of \; transition \; s \; from \; state \; s_i} \qquad [14]$$

$$b_{jk} = \frac{\exp\, ected \; number \; number \; of \; times \; in \; s_j and \; observatin g \; v_k}{\exp\, ected \; number \; of \; times \; in \, state \; j} \qquad [15]$$

Then, HMMs algorithms must to solve efficiently the problems mentioned above. For each state, the HMMs can use since one to five Gaussian mixtures both to reach high recognition rate and modelling vocal tract configuration in the Automatic Speech Recognition.

Gaussian mixtures

Gaussian Mixture Models are a type of density model which comprise a number of component functions, usually Gaussian. These component functions are combined to provide a multimodal density. They can be employed to model the colours of an object in order to perform tasks such as real-time colour-based tracking and segmentation. In speech recognition, the Gaussian mixture is of the form [Bilmes 98] [Resch, 2001a], [Resch, 2001b], [Kamakshi et al., 2002] and [Mermelstein, 1975].

:

$$g(\mu,\Sigma)(x) = \frac{1}{\sqrt{2\pi^d}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \qquad [16]$$

Equation 12 shows a set of Gaussian mixtures:

$$gm(x)=\sum_{k=1}^{K} w_k * g(\mu_k,\Sigma_k)(x) \qquad [17]$$

In 12, the summarize of the weights give us

$$\sum_{i=1}^{K} w_i = 1 \quad \forall \quad i \in \{1,\dots\dots\dots,K\} \quad : w_i \geq 0 \qquad [18]$$

Viterbi Trainning

We used Viterbi training, in this a set of training observations O^r, $1 \leq r \leq R$ is used to estimate the parameters of a single HMM by iteratively computing Viterbi alignments. When used to initialise a new HMM, the Viterbi segmentation is replaced by a uniform segmentation (i. e. each training observation is divided into N equal segments) for the first iteration.

Apart from the first iteration on a new model, each training sequence O is segmented using a state alignment procedure which results from maximising

$$\phi_N(T) = \max_i \phi_i(T) a_{iN} \qquad [19]$$

For *1<i<N* where

$$\phi_j(t) = \left[\max_i \phi_i(t-1) a_{ij}\right] b_j(o_t) \qquad [20]$$

With initial conditions given by

$$\begin{aligned} \phi_1(1) &= 1 \\ \phi_j(t) &= a_{1j} b_j(o_1) \end{aligned} \qquad [21]$$

For *i<j<N*. In this and all subsequent cases, the output probability $b_j(.)$ is as defined in the following equations:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=!}^{M_{js}} c_{jsm} \aleph(o_{st}; \mu_{jsm}, \sum{}_{jsm} \right]^{\gamma_s} \qquad [22]$$

If A_{ij} represents the total number of transitions from state i to state j in performing the above maximisations, then the transition probabilities can be estimated from the relative frequencies

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{k=2}^{N} A_{ik}} \qquad [23]$$

The sequence of states which maximises $\phi_N(T)$ implies an alignment of training data observations with states. Within each state, a further alignment of observations to mixture components is made.

We can use two methods for each state and each stream

1. use clustering to allocate each observation o_{st} with the mixture component with the highest probability
2. associate each observation o_{st} with the mixture component with the highest probability

In either case, the net result is that every observation is associated with a single unique mixture component. This association can be represented by the indicator function $\psi^{r}_{jsm}(t)$ which is 1 if o^{r}_{st} is associated with mixture component m of stream s of state j and zero otherwise.

The means and variances are then estimated via simple averages

$$\hat{\mu}_{jsm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi^{r}_{jsm}(t)o^{r}_{st}}{\sum_{r=!}^{R}\sum_{t=!}^{r}\psi^{r}_{jsm}(t)}$$

$$\hat{\sum}_{jsm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi^{r}_{jsm}(t)(o^{r}_{st} - \hat{\mu}_{jsm})(o^{r}_{st} - \hat{\mu}_{jsm})'}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{t=1}^{M_s}\psi^{r}_{jst}(t)} \qquad [24]$$

5 Experiments and Results

The evaluation of the experiment proposed involved 5 people (3 men and 2 women) with 300 speech sentences to recognize for each one. Speech signals were recorded at 11200 frequency sample, with 8 bits mono stereo (one channel) in laboratory environment, that significance that they were clean speech without noise. After that, speech signals were processed to eliminate information not useful that is the speech sentences were processed to find start and end points of speech signal, using software own. This software used the energy parameter to find it the interesting region.

Firstly, we used 1500 speech sentences extracted from 5 speakers individually (we used 100 for training task and 200 for recognition task), and we trained the Automatic Speech Recognition using Hidden Markov Models with 6 states (4 states with information and 2 dummies to connection with another chain). Also, we employed one Gaussian Mixture for each state in the chain Markov. The parameters extracted of the speech signal were 39 (13 MFCC, 13 delta and 13 energy coefficients), they are used to training the Hidden Markov Model.

The results obtained in this experiment are resumed in table 1.

Table 1 Recognition rate for 5 speakers individually

	users				
Recognition rate	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5
	99.5	99.5	98.5	98.52	98.5

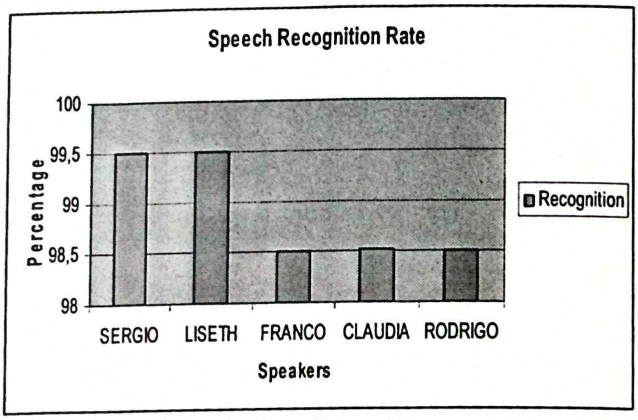

Fig 1 Graphical representation of the recognition rate reported in table 1

Secondly, we divided in two clusters our corpus, separating speech men of the women speech. The speech signals were the same that we used before. Finally, we integrated all speech signals (500 speech sentences, 100 for each speaker) in a corpus that we labelled as 'todos'. We trained the system newly and we probe this new corpus with the 1200 remaining. The results obtained in these two experiments are reported in table 2.

Table 2 Recognition rate for all, men and women speakers

	users		
Recognition rate	all	men	women
	94.13	95	98

Figure 2 shows a graphical representation of the results in table 2.

At respect, we obtain a 94.13% of successful recognition and 5.87% of error rate for all speakers. Given in this table are results for several speech sentences trained and recognized. The HTK (Hidden Markov Model Toolkit) was employed to obtain the results and it was based on complete words.

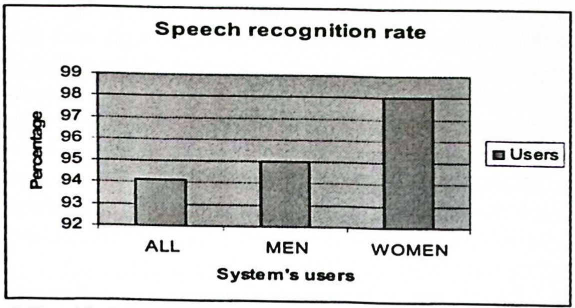

Fig 2 Graphical representation of the recognition rate reported in table 2

6 Conclusions and future works

The main purpose of this paper was to develop a fully ASR system using Hidden Markov Models. We included 5 people in our ASR system; it was because we constructed an ASR system independent of speaker. The results obtained demonstrated that ASR has a high performance independently of amount of speakers that it was included into it. Likewise, the Automatic Speech Recognition (ASR) for each speaker resulted to be satisfactory.

After the results obtained we revised the speech files that help us in this search. We must to say that we found problems with speech signals badly segmented in begin and end of the word. Overcoat for men speech files (especially third, fourth and fifth speakers). It obviously represented in bad results.

For future works we must to find an efficient algorithm that can split the speech signal in signal not necessary, useful signal and not useful signal. That is, we must to work in increase the number of the speakers and programming another splitting algorithm, and we will probably obtain better results. Though the results reported demonstrated a good performance.

References

[Bechetti and Prina 1999] Bechetti Claudio and Prina Ricoti Lucio, "Speech Recognition Theory and C++ Implementation", Fundazione Ugo Bordón, Rome, Italy, John Wiley and Sons, Ltd, 1999.

[Bilmes 98] BILMES J.A., "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", International Computer Science Institute, Berkeley, CA. 1998.

[DeFatta et al. 1988] DeFatta J. David, Lucas G. Joseph and Hodgkiss S. William, Digital Signal Processing, A system design approach, John Wiley & Sons, 1988.

[Farnetani 97] Farnetani E., "Coarticulation and connected speech processes", in the Handbook of Phonetic Sciences, W. Hardcastle and J. Laver, Eds., Blackwell, pp. 371-404 (1997).

[Kamakshi et al. 2002] KAMAKSHI V. Prasad, Nagarajan T. and Murthy Hema A. "Continuous Speech Recognition Using Automatically Segmented Data at Syllabic Units". Department of Computer Science and Engineering. Indian Institute of Technology. Madras, Chennai 600-636, 2002.

[Kirschning 1998] Kirschning Albers Ingrid, "Automatic Speech Recognition with the parallel Cacade Neural network", PhD Thesis, Tokyo Japan, March 1998.

[Mermelstein 1975] MERMELSTEIN Paul "Automatic Segmentation of Speech into Syllabic Units". Haskins Laboratories, New Haven, Connecticut 06510, pp. 880-883,58 (4), June 1975.

[Oppenheim 89] Oppenheim A. V., Shafer R. W., Digital Dignal Processing, Prentice Hall (1989).

[Rabiner and Biing-Hwang 1993] RABINER Lawrence and Biing-Hwang Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993.

[Resch 2001a] RESCH Barbara. "Gaussian Statistics and Unsupervised Learning". A tutorial for the Course Computational Intelligence Signal Processing and Speech Communication Laboratory. www.igi.turgaz.at/lehre/CI, November 15, 2001.

[Resch 2001b]. RESCH Barbara. "Hidden Markov Models". A Tutorial for the Course Computational Laboratory. Signal Processing and Speech Communication Laboratory. www.igi.turgaz.at/lehre/CI, November 15, 2001.

[Zhang 1999]. ZHANG Jialu, "On the syllable structures of Chinese relating to speech recognition", Institute of Acoustics, Academia Sinica Beijing, China, 1999.