Template-based and HMM-based Approaches for Isolated Spanish Digit Recognition

Lucas D. Terissi*, **Juan C. Gómez**

Laboratory for System Dynamics and Signal Processing
FCEIA, Universidad Nacional de Rosario
Riobamba 245 Bis, 2000 Rosario, Argentina
Phone: +54 341 4808543 - ext. 106
Fax: +54 341 4821772
Web: http://www.fceia.unr.edu.ar/lsd
lterissi@eie.fceia.unr.edu.ar, jcgomez@fceia.unr.edu.ar

Abstract

Isolated word recognition is usually performed by two different approaches, viz., Conventional Template-based and Hidden Markov Models (HMM)-based. A comparison between these approaches for isolated digit recognition (in Spanish language) is performed in this paper. For the template-based approach, several Dynamic Time Warping algorithms are proposed and implemented in a Matlab environment, while for the HMM-based approach, the algorithms available in a freeware toolbox has been employed. The algorithms are tested on a database collected during a Course on Digital Processing of Speech Signals at the University of Rosario. The test results show that a better performance can be obtained with a HMM-based digit recognizer.

Keywords: Isolated Word Recognition, Cepstral features, Hidden Markov Models, Dynamic Time Warping.

1 Introduction

Speech recognition by machine has been an area of intensive research worldwide during the last two decades. The impressive recognition performance obtained with the introduction of statistical models for speech, called Hidden Markov Models, has allowed the transfer of these techniques to many commercial applications ranging from voice activated access to banking and other services, through telephone voice dialing in mobile communications. Before the introduction of HMM in speech applications, the main approach for speech recognition was based on pattern comparison using reference templates in the cepstral domain. This template-based recognition approach has a more intuitive appeal, but the recognition performance is usually lower than that of the HMM-approach.

In Fig. 1, a block diagram representation of the classical pattern comparison approach for speech recognition is represented. The basic idea is to generate reference patterns for the words in the recognition vocabulary based on training data and then to compare them with the feature vector representing the word to be recognized. For the

*Author to whom all correspondence should be addressed.

case of the conventional template-based recognition approach, the reference patterns are deterministic models obtained from the feature vectors for each word in the training data [8]. In the case of HMM, the reference patterns are stochastic models given by the specification of three probability measures, namely, the state transition probability distribution (A), the observation symbol probability distribution (B) and the initial state distribution (π), which are obtained from the feature vectors for each word in the training data (for a detailed treatment on the use of HMM in speech recognition the reader is referred to [7],[2] and the references therein).

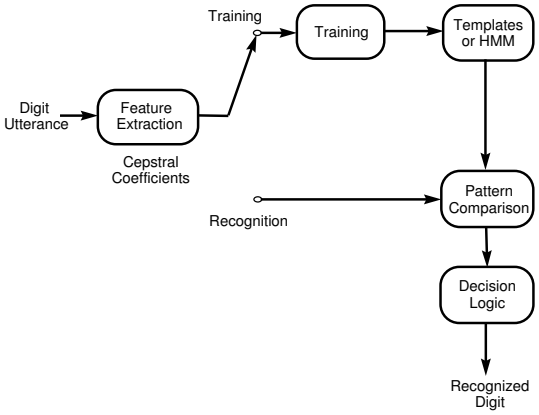

Figure 1: Block diagram of the pattern comparison approach for speech recognition.

In this paper, a comparison between these two approaches for isolated digit recognition is carried out. The proposed methods have been implemented in a Matlab environment [4], and they have been tested on a database generated from the Spanish utterances of the digits from zero to nine spoken by twenty five different speakers. The database was compiled during a course on Digital Processing of Speech Signals [6] at the University of Rosario (UNR). The test results show that the recognition rate is higher in the HMM approach with a lower computational load in comparison with the conventional template-based recognition approach.

2 Front End Processor

Feature extraction from speech samples is the first processing stage in the pattern comparison approach for both, template-based and HMM-based recognition. In Fig. 2, a block diagram representation of the processing tasks required for feature extraction is schematically depicted. The involved steps are as follows.

1. Pre-emphasis: speech signal is processed by a first order digital filter in order to spectrally flatten the signal.

2. Blocking into Frames: signal samples are grouped into length-N frames, and consecutive frames are overlapped in M signal samples.

3. Cepstral Analysis: The first L cepstral coefficients are computed for each frame conforming the feature vectors associated to each utterance.

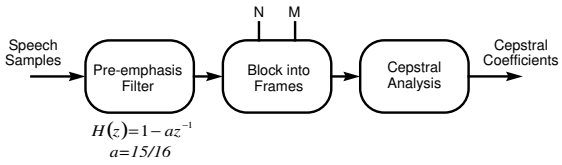

Figure 2: Block diagram of the front end processor for feature extraction.

3 Template-based Recognition Approach

In the training stage, feature vectors corresponding to the data samples are processed in order to generate a reference pattern vector for each digit. This is done by computing the centroid of the feature vectors associated to all the training occurrences of each digit. A problem associated with this computation comes from the fact that, in general, even for the same word, the acoustic realization of the word may vary significantly, for instance due to different articulatory rates.

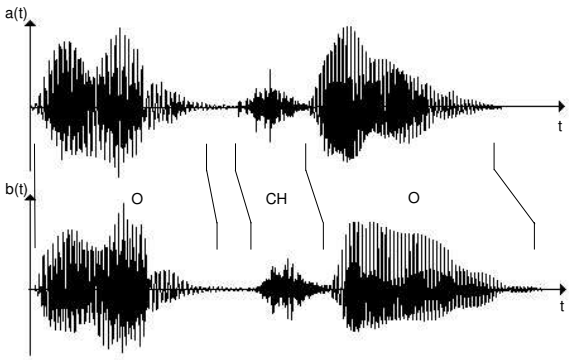

Figure 3: Two different utterances of the digit *ocho* (meaning *eight*). Segmentation between different phonemes is marked in solid lines.

Fig. 3 shows the waveforms corresponding to two different utterances of the Spanish digit *ocho* (meaning *eight*), where the need for time normalization between different utterances of the same word is apparent. It is clear that a linear time normalization is not appropriate and that a non-linear time alignment is required. This is typically solved in the literature using Dynamic Time Warping (DTW), a matching method that performs a non-linear time alignment to normalize the speaking rate fluctuations [2], [8].

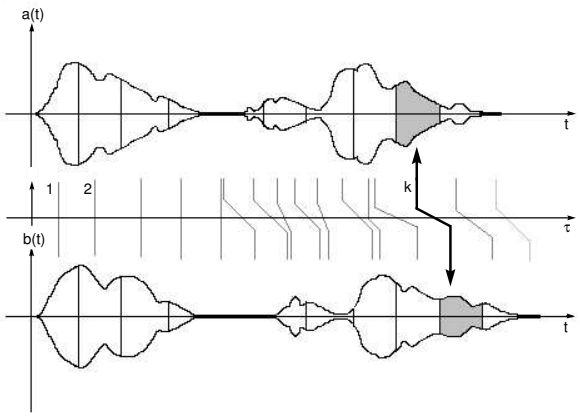

Figure 4: Dynamic Time Warping. Waveforms corresponding to two different utterances (a(t) and b(t)) associated to the digit *ocho* (meaning *eight*).

The basic idea of DTW is to non-linearly stretch or compress the signals in time, by referring the two speech patterns to a common *normal* time axis (τ), as can be seen in Fig. 4 for two different Spanish utterances corresponding to the digit *ocho* (meaning *eight*).

In the recognition stage, the feature vector associated to the digit to be recognized is compared to the reference pattern vectors, using a dissimilarity measure. DTW is performed previous to this comparison to time align the digit to be recognized with the templates. Then, the recognized digit is the one corresponding to the pattern vector with the smallest dissimilarity measure.

3.1 Template Generation

Feature vectors, for each digit, are generated as follows:

1. Choose one utterance from the training data, to use as a reference for the time-alignment.

2. Use the Dynamic Time Warping technique to align all the training data with the reference.

3. Once the training data are aligned, compute the reference pattern vector as the centroid of the feature vectors (of cepstral coefficients) corresponding to all the occurrences of the digit.

The choice of the reference will affect the performance of the recognition process. As there is no way to know which is the "best" reference, several alternatives can be proposed, among which the following can be mentioned.

- Take the digit sample of minimum duration.

- Take the digit sample of maximum duration.

- Take the digit sample whose duration is closest to the mean of the durations of all the training data.

- Take an arbitrary utterance of an arbitrary speaker for each digit.

4 Hidden Markov Model Approach

A system described by a Markov Chain (MC) is characterized by a finite number of distinct states, assumed to be observable, and a transition probability between states [3]. For speech applications these models are too restrictive to be of any practical utility. Hidden Markov Models extend the concept of MC to the case in which the states are only observable through a second stochastic process, resulting in a doubly embedded stochastic process [7]. An HMM (λ) is then parameterized by three probability measures, namely, the state transition probability distribution (A), the observation symbol probability distribution (B) and the initial state distribution (π). The notation $\lambda = (A, B, \pi)$ is usually employed.

There are several possible structures for HMM, but for speech applications the preferred one is the *left-right model* since the time can be associated with the states in a straightforward manner, and as the time increases the states proceeds from left to right [8]. An schematic representation of a 5-state left-right HMM is shown in Fig. 5.

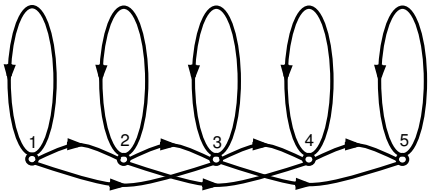

Figure 5: A 5-state left-right HMM structure.

For the case considered in this paper, the vocabulary to be recognized is composed by the ten Spanish utterances of the digits from zero to nine, where each digit is represented by a distinct HMM. Available for the recognition, there is a training data set composed by three different occurrences of each digit spoken by twenty different individuals. In the training stage, each utterance is converted to the cepstral domain and constitutes an observation sequence for the estimation of the HMM parameters associated to the respective digit. The estimation is performed by optimizing the likelihood of the training vectors corresponding to each digit in the vocabulary. Typically, the optimization is performed using the Baum-Welch algorithm or equivalently the EM (Expectation-Maximization) algorithm [1]. In the recognition stage, the observation sequence representing the digit to be recognized is used to compute the likelihoods, for all possible models, that the sequence has been generated by these models. The recognized digit corresponds to the one associated to the model with the highest likelihood. In this stage the Viterbi algorithm, as described in [9], is employed.

Fig. 6 schematically represents an isolated digit recognizer based on HMMs.

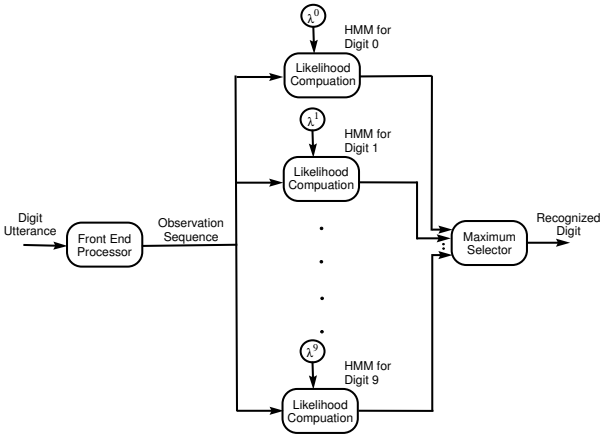

Figure 6: Block diagram of an isolated digit HMM recognizer (adapted from [7]).

5 Performance Comparison

In this section, a performance comparison of the conventional template-based and the HMM-based recognition approaches is carried out.

The database (digitized at a 11.025kHz rate), in Spanish language, was generated from 25 speakers (13 males and 12 females) with 3 different occurrences of each digit spoken by each talker (students attending a course on Digital Processing of Speech Signals at the UNR [6]). The digits were segmented and the background noise was suppressed via a spectral substraction. For the training stage a set of 20 speakers (10 males and 10 females) was used. To test the performance of both recognition algorithms when utterances of speakers not belonging to the training set are being recognized, the remaining set of 5 speakers, considered as external talkers, was used. All the algorithms were programmed in the Matlab environment [4].

Table 1: Average error rate (AER) for each digit (Dj, j = 0, 1, ··· , 9) for the conventional template-based recognition approach, using each speaker (Si) as a reference for the time-alignment.

Ref.	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	AER
S1	8.33	13.89	2.78	5.55	27.78	0.00	13.89	2.78	2.78	2.78	8.06
S2	13.89	2.78	0.00	2.78	33.33	0.00	8.33	8.33	12.22	11.11	10.28
S6	16.66	0.00	2.78	0.00	16.66	0.00	16.66	0.00	5.56	5.56	6.39
S8	5.55	0.00	0.00	0.00	33.33	0.00	13.89	11.11	2.78	33.33	10.00
S11	5.56	2.78	0.00	0.00	33.33	5.56	5.56	8.33	16.66	8.33	8.61
S12	0.00	0.00	5.55	0.00	33.33	0.00	11.11	5.55	5.55	8.33	6.94
S13	27.78	2.78	0.00	0.00	30.55	0.00	25.00	2.78	44.44	5.55	13.89
S15	13.89	8.33	0.00	0.00	27.78	2.78	5.55	2.78	5.55	2.78	6.95
S18	8.33	0.00	0.00	0.00	36.11	0.00	11.11	5.55	5.55	27.78	9.44
S20	16.67	0.00	0.00	0.00	38.89	2.78	22.22	16.66	36.11	11.11	14.44

For the conventional template-based approach, at the front end processor, signal samples were grouped into 20ms-length frames, consecutive frames were overlapped in 13ms and the first nine cepstral coefficients were computed for each frame to conform the feature vectors associated to each utterance. Codebooks were generated using the different reference selection criteria described in subsection 3.1. Table 1 shows the Average Error Rate (AER) for each digit when the utterances of each speaker are used as a reference for the time-alignment. Here, speakers belonging to the training set were considered to compute the AER.

Table 2: Average error rate (AER) for each digit (Dj, j = 0, 1, ··· , 9) for the conventional template-based recognition approach, using the longest (L), the shortest (S), and the mean (M) utterance durations, as a reference for the time-alignment.

Dur.	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	AER
L	21.67	5.00	5.00	6.33	38.33	0.00	45.00	0.00	15.00	18.33	15.5
S	15.00	0.00	0.00	0.00	36.67	0.00	21.67	11.67	18.33	16.67	12.00
M	13.89	0.00	0.00	0.00	30.55	2.78	11.11	2.78	5.55	2.78	6.95

Table 3: Average error rate (AER) for each digit (Dj, j = 0, 1, ··· , 9) for the HMM-based recognition approach with 2, 3, 4 and 5 state left-right models.

# states	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	AER
2	1.67	0.00	0.00	5.00	6.67	0.00	3.33	0.00	1.67	1.67	2.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.02
4	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.03
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Comparison of the Average error rate (AER) for each digit (Dj, j = 0, 1, ··· , 9) between HMM-based and Conventional Template-Based (CTB) recognition approaches with external speakers.

Approach	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	AER
CTB	26.67	20.00	0.00	6.67	33.33	0.00	20.00	20.00	20.00	40.00	18.67
HMM	0.00	13.33	6.67	6.67	0.00	0.00	0.00	6.67	0.17	26.67	6.00

Table 2 shows the average error rate for each digit for three different reference alignment selection criteria, namely, the utterance of shortest duration, the utterance of largest duration, and the utterance with duration closest to the mean of all digit durations.

The same database was used to estimate a HMM for each digit. At the front end processor signal samples were grouped into 20ms-length frames and the first nine cepstral coefficients were computed for each frame to conform the feature vectors associated to each utterance, no overlapping were used in this approach. The HMM Module of the Bayes Net Toolbox for Matlab by Murphy [5] was employed for the simulations. Tests were performed for model structures with two, three, four and five states. The best performance was obtained for a 5-state model structure, so this model structure was the one selected for the comparison with the conventional template-based approach. Table 3 shows the average error rate for all digits for 2, 3, 4 and 5-state left-right HMM structures.

As can be observed from Tables 1 and 2, the best performance for the conventional template-based recognition approach is obtained using the utterances with duration closest to the mean of all digit durations as the reference for the time-alignment. This result is then compared with the corresponding to the HMM-based approach shown in Table 3. As can be observed, the HMM-based recognition approach outperforms the conventional template-based one.

Recognition tests using external speakers were also carried out. The results are shown in Table 4, where the average error rate for each digit for both approaches when the set of external speakers' digit utterances is being recognized are presented. The recognition rates shown in the table correspond to the models of better performance with the training data, *i.e.*, the template model with time-alignment using the mean of all digit durations as a reference for the Conventional Template-Based (CTB) approach, and the 5-state left-right model, for the HMM approach.

The tests, both with training data and external speakers, show that the HMM recognition approach outperforms the CTB recognition approach. An extension of the database is currently under development to include a larger number of speakers, in order to improve the robustness of the algorithms against speaker variability.

6 Conclusions

In this paper, a comparison between conventional template-based and HMM-based approaches for isolated digit recognition (in Spanish language) has been performed. For the CTB approach, several algorithms for Dynamic Time Warping have been implemented in a Matlab environment, while for the HMM-based approach the Toolbox in [5] was employed. The performance comparison tests were carried out on a database compiled during a course on Digital Processing of Speech Signals at the UNR. The test results show that the HMM-based recognition approach outperforms the conventional template-based approach, even for a relatively small database.

References

[1] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistics functions of Markov chains. *Annals of Mathematics and Statistics*, 41:164–171, 1970.

[2] X. Huang, Y. Ariki, and M. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[3] D. Kannan. *An Introduction to Stochastic Processes.* North Holland, New York, 1979.

[4] The MathWorks. Developers of MATLAB, Simulink, and Stateflow for technical computing. http://www.mathworks.com/, 2002.

[5] K. Murphy. Hidden Markov Models module in Bayes net toolbox for Matlab. http://www.ai.mit.edu/∼murphyk, 2004.

[6] ProDiVoz. Digital processing of speech signals (Procesamiento Digital de Señales de Voz). http://www.fceia.edu.ar/prodivoz, 2004.

[7] L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[8] L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, New Jersey, 1993.

[9] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, IT-13:260–269, 1967.