ISSN 1870-4069

Digit-Based Speaker Verification in Spanish Using Hidden Markov Models

Juan Carlos Atenco Vazquez[1], Juan Carlos Moreno Rodriguez[1], Rene Arechiga Martinez[2], Juan Manuel Ramirez Cortes[1], Pilar Gomez Gil[1], Rigoberto Fonseca Delgado[3]

[1] National Institute of Astrophysics, Optics and Electronics, Mexico

[2] New Mexico Tech., Electrical Engineering Department, USA

[3] Yachai Tech., Ecuador

{atencovaz, jmram, jmram, pgomez}@inaoep.mx, rene.arechiga@nmt.edu, rfonseca@yachaytech.edu.ec

Abstract. In this paper we propose a digit-based text-dependent speaker verification system (SVS) in Spanish. The system uses word level Hidden Markov Models (HMM) as classifiers and Frequency Cepstral Coefficients (MFCC) with Cepstral Mean Subtraction as features. The system was developed considering a gender independent Universal Background Model (UBM) within an HMM-UBM framework. The training data set was pooled with both genders equally represented for the UBM. A Target Speaker Model (TSM) was generated using Maximum A Posteriori (MAP) to adapt the UBM's parameters to the acoustic characteristics of the target speakers. Every target speaker (TS) has a 4 digits password. Robustness of the system was tested using adaptive noise cancellation as a speech enhancement scheme; the speech signals were corrupted with additive white Gaussian noise (AWGN) at different values of signal to noise ratio. Generation of a digit-database, which we have named as BIOMEX-DB, is described. The main contribution of this work is a robust SVS in Spanish language tested with 4-digits passwords, which can be easily adapted to different lengths or text-prompted mode. Obtained results showed an equal error rate (EER) in the range of 1.0567-1.4465 % when 50 subjects in the database were considered.

Keywords: speaker verification, HMM, voice biometrics, universal background model, speaker adaptation, voice database.

1 Introduction

Nowadays there is an increasing number of applications that require verification of a user's identity, such as access to facilities, internet applications, or bank

services. In that sense, biometric systems are widely used as a viable solution for most security problems.

A biometric modality with a good balance of properties within the universe of desired conditions is the use of voice signals, known as speaker verification, which constitutes an active research area. Many approaches focused on attacking the inherent problems of this biometric modality have been recently proposed [11]. In a speaker verification task there are two relevant modalities: text dependent and text independent. In the first scheme, the words spoken by a speaker for recognition are limited to a specific vocabulary, in the second approximation there is no limitation in the words that can be uttered [17]. There is an increasing interest to capitalize on the advantages of text-dependent systems while allowing for the flexibility of the text-independent domain [4]. A review of text dependent modality can be consulted in [6].

HMM and Gaussian Mixed Models (GMM) have been widely used in speaker recognition approaches. In recent years these techniques have been used to obtain statistics for a new representation known as i-vector feature extraction [19, 20]. In [16] HMM's are used for data segmentation and the resulting statistics are incorporated in a system based on Joint Factor Analysis (JFA). In other works these models have been used as classifiers aiming to capture the temporal information of voice signals and improve accuracy through different types of UBMs [14, 13]. In [7] a performance comparison is made between HMM, GMM and i-vector showing competent results in different scenarios. Another important aspect of speaker recognition is the choice of a feature extraction technique that improves accuracy and captures the information of the speech signal. Several techniques for representing spectral features, such as Linear predictive cepstral coefficients (LPCC), Perceptual Linear Predictive (PLP) and Mel Frequency Cepstral Coefficients (MFCC) have been used for different speech processing tasks, mainly speech and speaker recognition. Reference [8] presents details on the use of LPC and MFCC with several classifiers for speech recognition. In [3] the performance of a HMM based speech recognition system trained with noisy speech samples using several spectral feature extraction techniques is compared. In the field of speaker recognition, spectral features have a widespread use, in [1] spectral features are used to train several classification models for a text independent identification task.

In this work, a HMM-based speaker verification system using a single 4 digits password in Spanish is presented. Verification is done through the matching of the spectral characteristics of the voice using Universal Background Models (UBM) and Target Speaker Models (TSM), obtaining a score based on Log Likelihood Ratio (LLR), testing speaker correct (SC) and Impostor correct (IC) for the assessment of identity verification. The MFCC feature extraction technique was used together with the Delta (first derivative) and Delta-Delta (second derivative) coefficients. Cepstral Mean Normalization (CMN) technique was applied to compensate for spectral effects caused by the recording channel. Each HMM is trained at word level representing a digit, so there is a set of UBMs composed of the 10 models plus 1 silence model, and a target-speaker set

using the same scheme. Robustness of the system in several noise conditions was tested using an adaptive LMS filter approach.

Results are analyzed using ROC (Receiver Operating Characteristic) curves, presenting an equal error rate (EER) in the range of 1.0567-1.4465 % with a population of 50 subjects. The rest of the paper is organized as follows: section 2 describes theoretical concepts; section 3 presents a description of the BIOMEX-DB database which was generated for this experiment. In section 4, experimental development and performed testing on the SVS are explained. Tests results are described in section 5 and concluding remarks are presented in section 6.

2 Speaker Verification System

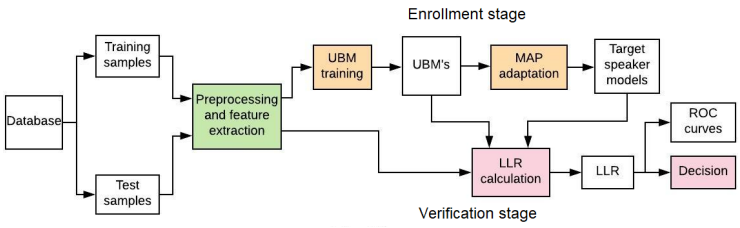

Fig. 1. Speaker verification process.

Figure 1 shows the block diagram of the SVS indicating the enrollment and verification stages. Feature extraction consists of the generation of MFCC [5], incorporating Delta and Delta-Delta in the feature vector. The Enrollment stage consists of training of HMM-based UBMs using the Baum-Welch algorithm for parameter estimation[18], and after a MAP adaptation TSMs are generated. In the verification stage, LLR is calculated to qualify the match between the features of a test speech sample and both UBM and TSM [12]. Log likelihood ratio is defined in Equation 1.

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_S) - \log p(\mathbf{X}|\lambda_{UBM}), \tag{1}$$

where $p(\boldsymbol{X}|\lambda_S)$ is the likelihood that the feature vector $\mathbf{X}$ has been generated by the speaker model λ_S, while $p(\boldsymbol{X}|\lambda_{UBM})$ is the likelihood that $\mathbf{X}$ was generated by the UBM. Verification is then carried out by comparing the LLR with a predetermined threshold and based on that comparison the SVS accepts or rejects the claimed identity of a test speaker.

2.1 Hidden Markov Models (HMM)

Figure 2 shows a block diagram of the HMM structure used in this work. An HMM represents a doubly embedded stochastic dynamical process through a set

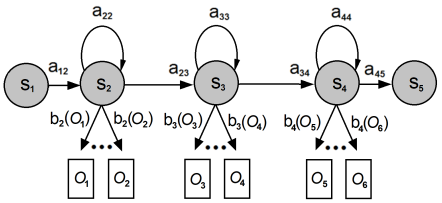

Fig. 2. Hidden Markov Model (HMM) representation.

of states $S : \{S_1, S_1, ..., S_N\}$, a set of possible observations $O : \{O_1, O_1, ..., O_M\}$, a transition probability matrix $A = [A_{ij}]$, and an output probability distribution $B = [B_{ij}]$. The stochastic nature of HMMs is contained in the transition of states defined in the transition matrix A, and in generation of the observations B, being both probabilistic events [10]. HMM have been widely used in speech processing due to the ability to capture temporal information of the speech, taking into account its non-stationary nature. In this work, the HTK (Hidden Markov Model Toolkit) software tools [18] were used to build the UBMs and TSMs conforming the SVS. Parameter estimation and frame alignment in HMM were implemented using Baum-Welch and Viterbi algorithms respectively [18].

2.2 Speaker Adaptation

A TSM contains the acoustic characteristics of a TS. The process followed in this work was the adaptation of the parameters of previously trained UBMs with the target speaker's speech information. MAP parameter adaptation is defined according to equation 2:

$$\widehat{\mathbf{u}}_{jm} = \frac{\mathcal{N}_{jm}}{\mathcal{N}_{jm} + \tau}\overline{\mathbf{u}}_{jm} + \frac{\tau}{\mathcal{N}_{jm} + \tau}\mathbf{u}_{jm}, \tag{2}$$

where $\widehat{\mathbf{u}}_{jm}$ is the adapted mean vector, $\overline{\mathbf{u}}_{jm}$ is the mean vector of the adaptation data and $\mathbf{u}_{jm}$ is the mean vector of the speaker independent model. $\mathcal{N}_{jm}$ is the likelihood that the adaptation data was generated by the Gaussian component m in the j state, while τ is the weight of the adaptation data. In [2] it is shown that adapting other parameters, mean not included, decreases the accuracy of an SVS, furthermore, it is also shown that a very high value of τ can also decrease the accuracy. In this work, the TSMs were generated by adapting only the means, with $\tau = 12$ established by experimentation.

3 BIOMEX-DB Database Generation

The experiments described in this work were carried out with a database created specifically for biometric purposes, focused in digit-based text dependent speaker verification in Spanish language. The speech signals were obtained from 51 volunteers, 26 men and 25 women to ensure a balanced representation of both

genders and voice variability. Table 1 shows the demographic distribution of volunteers within the database.

Table 1. Demographic distribution of subjects in the database.

	Gender	
Age	Male	Female
Less than 21	3	2
21 - 30	18	14
31 - 40	2	6
41 - 50	1	1
51 - 60	0	2
Older than 60	2	0

The speech database consists of audio files containing strings of digits randomly ordered with a short segment of silence between utterances. The speech database is divided into two parts: The first part consists of 10 strings of 10 digits each, giving a total number of 5100 digits pronounced by 51 subjects. The second part consists of 10 strings of 4 digits each. Each string is considered a 4 digits password assigned to a specific speaker. Similarly, each digit is pronounced once per string. All speech samples of each speaker were recorded consecutively in 15-20 minutes sessions. The recording was carried out with a microphone Sennheiser model MD 421-II connected to a deskstop computer through a Yamaha amplifier model MG06X. A MATLAB script was used to display the digits to be pronounced and to generate the transcripts of each audio file. The audio signals were recorded with a sampling frequency of 16 KHz, a resolution of 16 bits per sample, and stored in wav format.

4 Experimental Setup

The evaluation was conducted using two gender balanced population sizes of 40 and 50 speakers. These population sizes were chosen according to the maximum amount of speakers available in the database and to evaluate the impact of different population sizes in the results. With a population of 40 speakers, 24 were employed to train the UBM, 8 target speakers and 8 impostors. With 50 speakers, 30 were employed to train the UBM, 10 target speakers and 10 impostors, using gender balance in all cases. The MFCC features were extracted using signal framing of 25 ms Hamming windows with an overlapping of 10 ms. The number of states of the HMM was the same in every UBM and TSM models, with one Gaussian component per state. Three cases were analyzed according to the number of states in the HMM structure: 5, 8 and 12 states. Testing was carried out using a jack-knife-like iterative scheme [9].

Robustness of the system in the presence of noise was tested through a series of experiments in which the speech signals were corrupted with additive white

Gaussian noise at different signal-to-noise values. For that purpose, a normalized least mean square (NLMS) filter was incorporated [15].

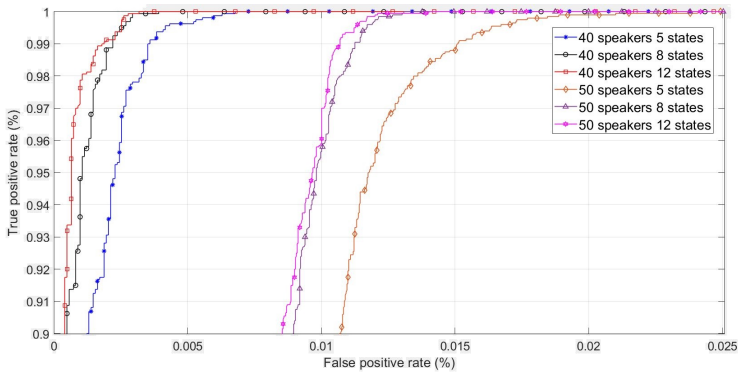

Fig. 3. Speaker Verification results; ROC curves.

5 Results and Discussion

Fig. 3 shows the ROC curves obtained with two population cases, 40 and 50 subjects, and three different HMM models with 5, 8 and 12 states. This figure has been zoomed to the upper left corner in order to highlight details. These results are concentrated in table 2 showing the evaluation parameters EER and Area under the Curve (AUC).

The obtained EER is located in the ranges 0.2516-0.4353 % and 1.0567-1.4465 % for the cases of 40 and 50 speakers, respectively. Results obtained in the proposed work are comparable with those presented in recent works, although differences in population sizes, corpora, training conditions and language do not allow a direct evaluation. In [13] the authors report HMM-UBM and GMM-UBM systems with an EER ranging from 7.12 to 0.79 %.

Reference [7] presents different HMM and GMM systems with an EER in the range 9.94-0.59 %. In [14], HMM-UBM and GMM-UBM systems with an EER of 5.56-0.009 % are reported. Results show that the more states the HMMs have the better the verification results are, since there are more states the acoustic modeling improves, however 12 states HMMs don't show significant improvements in comparison with 8 states because there isn't enough training data to further improve the results.

A trade off between the computational cost for both training and testing and the number of states must be take into consideration, more states increases the computational cost, therefore the 8 states HMMs were considered good choices. It is evident that the best results come from the 40 subjects population, it is well documented that performance decreases with big populations.

Table 2. Speaker verification results; EER and AUC.

	Population			
HMM states	P=40		P=50	
N	EER %	AUC	EER %	AUC
5	0.4353	0.9995	1.4465	0.9943
8	0.2992	0.9998	1.1378	0.9951
12	0.2516	0.9999	1.0567	0.9952

Table 3 shows the values of EER and AUC obtained when the speech signals are corrupted with AWGN at the following SNR values: -10 dB, -5 DB, and 0 dB, using a HMM structure of 8 states. As expected, the system performance decreases compared to the noiseless case, however, the obtained EER lies in the range of 0.7467-0.8746 % which it still a satisfactory system behavior.

Table 3. EER and AUC results with signals corrupted by AWGN; N=8 HMM states, P=40 subjects.

SNR (dB)	EER (%)	AUC
Noiseless	0.2992	0.9999
-10	0.7467	0.9997
-5	0.8107	0.9997
0	0.8746	0.9997

6 Conclusions

A HMM-UBM based speaker verification system using a four-digit password pronounced in Spanish language, has been presented. Several experiments were carried out with different number of emitting states and two population sizes: 40 and 50 speakers. The more emitting states the better obtained accuracy, however, the training time increases and more computational resources are needed.

The experiments showed that the best results in terms of EER, AUC, and computational cost considerations were obtained with a HMM structure of 8 states. This structure was further tested in noise conditions adding white Gaussian noise at several SNR values, showing good noise tolerance. Results showed a degradation on EER from 0.2992 % without noise to 0.8746 % with a SNR=0 dB which corresponds to a noise power level in the same magnitude of the voice signal. In practical applications is not possible to operate in noiseless environments, so adaptive filtering is an affordable alternative to improve the robustness of a HMM-UBM based SVS. Additional experiments with real and typical noise conditions are currently in progress. In conclusion, the system showed high accuracy to discriminate between correct target speakers and impostors.

References

1. Charan, R., Manisha, A., Karthik, R., Kumar, M.R.: A text-independent speaker verification model: A comparative analysis. In: 2017 International Conference on Intelligent Computing and Control (I2C2). pp. 1–6. IEEE (2017)
2. Gauvain, J., Barras, C.: Feature and score normalization for speaker verification of cellular data. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP'03). vol. 2, pp. II–49. IEEE (2003)
3. Këpuska, V.Z., Elharati, H.A.: Robust speech recognition system using conventional and hybrid features of mfcc, lpcc, plp, rasta-plp and hidden markov model classifier in noisy conditions. Journal of Computer and Communications 3(06), 1 (2015)
4. Kofi, B.: Speaker recognition in the text-independent domain using keyword hidden markov models. Masters Report, University of California at Berkeley (2005)
5. Kopparapu, S.K., Laxminarayana, M.: Choice of mel filter bank in computing mfcc of a resampled speech. In: 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010). pp. 121–124. IEEE (2010)
6. Larcher, A., Lee, K.A., Ma, B., Li, H.: Text-dependent speaker verification: Classifiers, databases and rsr2015. Speech Communication 60, 56–77 (2014)
7. Liu, Y., He, L., Tian, Y., Chen, Z., Liu, J., Johnson, M.T.: Comparison of multiple features and modeling methods for text-dependent speaker verification. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 629–636. IEEE (2017)
8. Madan, A., Gupta, D.: Speech feature extraction and classification: A comparative review. International Journal of computer applications 90(9) (2014)
9. Martın-Donas, J.M., López-Espejo, I., González-Lao, C.R., Gallardo-Jiménez, D., Gomez, A.M., Pérez-Córdoba, J.L., Sánchez, V., Morales-Cordovilla, J.A., Peinado, A.M.: Secuvoice: A spanish speech corpus for secure applications with smartphones (2016)
10. Nguyen, L.: Tutorial on hidden markov model. Applied and Computational Mathematics 6(4-1), 16–38 (2017)
11. Poddar, A., Sahidullah, M., Saha, G.: Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics 7(2), 91–101 (2017)
12. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital signal processing 10(1-3), 19–41 (2000)
13. Sarkar, A.K., Tan, Z.H.: Text dependent speaker verification using un-supervised hmm-ubm and temporal gmm-ubm. In: interspeech. pp. 425–429 (2016)
14. Sarkar, A.K., Tan, Z.H.: Incorporating pass-phrase dependent background models for text-dependent speaker verification. Computer Speech & Language 47, 259–271 (2018)
15. Sil, R., Bharath, K., Karthik, R., Kumar, M.R.: Nlms-loess algorithm for adaptive noise cancelation. In: Microelectronics, Electromagnetics and Telecommunications, pp. 65–74. Springer (2019)
16. Stafylakis, T., Alam, M.J., Kenny, P.: Text-dependent speaker recognition with random digit strings. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24(7), 1194–1203 (2016)
17. Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S., Wang, R.: Speaker identification features extraction methods: A systematic review. Expert Systems with Applications 90, 250–271 (2017)

18. Young, S., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The htk book version 3.4 manual. Cambridge University Engineering Department, Cambridge, UK (2006)
19. Zeinali, H., Sameti, H., Burget, L., Cernockỳ, J., Maghsoodi, N., Matejka, P.: i-vector/hmm based text-dependent speaker verification system for reddots challenge. In: InterSpeech. pp. 440–444 (2016)
20. Zeinali, H., Sameti, H., Burget, L., et al.: Text-dependent speaker verification based on i-vectors, neural networks and hidden markov models. Computer Speech & Language 46, 53–71 (2017)