A Case Study of Speech Recognition in Spanish: from Conventional to Deep Approach

Aldonso Becerra, *Student Member, IEEE*, J. Ismael de la Rosa, *Senior Member, IEEE*, and Efrén González
Unidad Académica de Ingeniería Eléctrica
Universidad Autónoma de Zacatecas
Av. López Velarde No. 801, Col. Centro, C.P. 98068, Zacatecas, México
Email: a7donso@uaz.edu.mx, ismaelrv@ieee.org, gonzalez_efren@hotmail.com

***Abstract*—The aim of this paper is to exhibit a comparative case study of the conventional speech recognition GMM-HMM (Gaussian mixture model - hidden Markov model) architecture and the recent model based on deep neural networks. During years the GMM approach has controlled the speech recognition tasks, however it has been surpassed with the resurgence of artificial neural networks. To exemplify these acoustic modeling frameworks, a case study has been conducted by using the Kaldi toolkit, employing a personalized speaker-independent mid-vocabulary voice corpus for recognition of digit strings and personal name lists in latin spanish on a connected-words phone dialing task. The speech recognition accuracy obtained in the results shows a better word error rate by using the DNN acoustic modeling. A 20.71% relative improvement is obtained with DNN-HMM models (3.33% WER) in respect to the lowest GMM-HMM rate (4.20% WER).**

I. Introduction

Automatic speech recognition (ASR) systems are computer programs focused to produce a transcription from an input speech signal. GMM-HMM (Gaussian mixture model-hidden Markov model) approach has been used in speech recognition during years [1], [2], and wherein the dynamics of speech signal is modeled by HMMs, and the concordance of input observations to an HMM state is defined by GMMs. These models are suitable for generating the emission probabilities in each HMM state, but GMMs have a serious deficiency: they are inefficient for modeling data that lie on or near a nonlinear manifold in the data space [3]. An alternative way to perform the GMMs function is to employ deep neural networks (DNNs). This approach has the potential to learn better models of data that alleviate this shortcoming [4], originating the so-called DNN-HMM framework.

Acoustic modeling with artificial neural networks has shown it can outperform GMMs in speech recognition tasks [3], [5]–[8]. Even though two decades ago one-layer neural networks were employed in speech recognition tasks [9], neither the hardware nor the training algorithms were suitable on large amounts of data and complex network architectures. Recently, advances in machine learning and computer hardware make possible to offset those adversities [10]. DNNs modeling is carried out by using a two-stage training [8]: i) initialization of the weights (with or without pretraining) and ii) discriminative back-propagation training (fine-tuning). Pretraining, an alternative way to conventional random initialization of the weights, can be performed by some strategies like discriminative pretraining (DPT) [5], deep belief nets (DBN) [7], [11], autoencoders, dropout [12] and hybrid pretraining [13].

This paper compares two ASR architectures applied to a case study in spanish: a) conventional approach of GMM-HMM and b) nonlinear model of DNN-HMM. The case study has been implemented by using the Kaldi toolkit [14]. The process has been produced with a mid-vocabulary corpus on a personalized task of speaker-independent connected words in latin spanish. The results of DNNs model (3.33% WER) achieve a relative word error rate reduction of 20.71% in comparison with the lowest GMMs result (4.20% WER).

The rest of the paper is organized as follows. Section 2 briefly describes the ASR approaches. Some comparative experiments are shown in section 3. Finally, section 4 concludes the paper with a discussion of the work.

II. Speech Recognition Baseline

A. ASR Principles

In order to recognize the input speech, this must be represented through feature vector sequences called *acoustic observations*, $O = \{\vec{o}_1, \ldots, \vec{o}_T\}$, where $\vec{o}_t$ is the speech feature vector at time t. Given an acoustic observation O, the objective is to find the corresponding word sequence $W = w_1, \ldots, w_M$. ASR is usually stated using the Bayes Rule:

$$\hat{W} = \underset{w}{\operatorname{argmax}}\ p(W|O) = \underset{w}{\operatorname{argmax}}\ \frac{p(W)p(O|W)}{p(O)}, \quad (1)$$

where $\hat{W}$ is the most likely word sequence given O; observation O is usually a fixed value, thus the term $p(O)$ is omitted. The term $p(O|W)$ is called *acoustic model* (AM), and the term $p(W)$ is known as *language model* (LM) [1], [2].

B. GMM-HMM Approach

This architecture is parameterized by $\lambda = (A, \pi, B)$ [15], where $A \triangleq a_{ij}$ represents the HMM-states transition matrix, π is a probability distribution of initial HMM-state; $B = \{b_1(o_t), \ldots, b_j(o_t)\}$, where $b_j(o_t) \triangleq \sum_{i=1}^{M} w_i N(o_t|\mu_i, \Sigma_i)$, and w_i are the mixture weights, μ_i is the mean vector and Σ_i is the covariance matrix of the Gaussian N, and corresponds to the GMM of state j, $p(o_t|q_j)$, which denotes an emission probability of the input observation o_t given a state q_j at time t. The AM parameters λ can be efficiently estimated (with maximum

978-1-5090-2532-9/16/$31.00 ©2016 IEEE