World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:06, 2018

Semi-Supervised Learning for Spanish Speech Recognition Using Deep Neural Networks

Authors : B. R. Campomanes-Alvarez, P. Quiros, B. Fernandez

Abstract : Automatic Speech Recognition (ASR) is a machine-based process of decoding and transcribing oral speech. A typical ASR system receives acoustic input from a speaker or an audio file, analyzes it using algorithms, and produces an output in the form of a text. Some speech recognition systems use Hidden Markov Models (HMMs) to deal with the temporal variability of speech and Gaussian Mixture Models (GMMs) to determine how well each state of each HMM fits a short window of frames of coefficients that represents the acoustic input. Another way to evaluate the fit is to use a feed-forward neural network that takes several frames of coefficients as input and produces posterior probabilities over HMM states as output. Deep neural networks (DNNs) that have many hidden layers and are trained using new methods have been shown to outperform GMMs on a variety of speech recognition systems. Acoustic models for state-of-the-art ASR systems are usually training on massive amounts of data. However, audio files with their corresponding transcriptions can be difficult to obtain, especially in the Spanish language. Hence, in the case of these low-resource scenarios, building an ASR model is considered as a complex task due to the lack of labeled data, resulting in an under-trained system. Semi-supervised learning approaches arise as necessary tasks given the high cost of transcribing audio data. The main goal of this proposal is to develop a procedure based on acoustic semi-supervised learning for Spanish ASR systems by using DNNs. This semi-supervised learning approach consists of: (a) Training a seed ASR model with a DNN using a set of audios and their respective transcriptions. A DNN with a one-hidden-layer network was initialized; increasing the number of hidden layers in training, to a five. A refinement, which consisted of the weight matrix plus bias term and a Stochastic Gradient Descent (SGD) training were also performed. The objective function was the cross-entropy criterion. (b) Decoding/testing a set of unlabeled data with the obtained seed model. (c) Selecting a suitable subset of the validated data to retrain the seed model, thereby improving its performance on the target test set. To choose the most precise transcriptions, three confidence scores or metrics, regarding the lattice concept (based on the graph cost, the acoustic cost and a combination of both), was performed as selection technique. The performance of the ASR system will be calculated by means of the Word Error Rate (WER). The test dataset was renewed in order to extract the new transcriptions added to the training dataset. Some experiments were carried out in order to select the best ASR results. A comparison between a GMM-based model without retraining and the DNN proposed system was also made under the same conditions. Results showed that the semi-supervised ASR-model based on DNNs outperformed the GMM-model, in terms of WER, in all tested cases. The best result obtained an improvement of 6% relative WER. Hence, these promising results suggest that the proposed technique could be suitable for building ASR models in low-resource environments.

Keywords : automatic speech recognition, deep neural networks, machine learning, semi-supervised learning
Conference Title : ICMLA 2018 : International Conference on Machine Learning and Applications
Conference Location : Copenhagen, Denmark
Conference Dates : June 11-12, 2018