Speech Recognition using Deep Neural Networks Trained with Non-uniform Frame-Level Cost Functions

Aldonso Becerra[1], *Student Member, IEEE,*
J. Ismael de la Rosa[2], *Senior Member, IEEE,*
Efrén González[3], A. David Pedroza[4],
and J. Manuel Martínez[5]
Unidad Académica de Ingeniería Eléctrica
Universidad Autónoma de Zacatecas
Av. López Velarde No. 801,
Col. Centro, C.P. 98068, Zacatecas, México
Email: [1]a7donso@uaz.edu.mx, [2]ismaelrv@ieee.org,
[3]gonzalez_efren@hotmail.com, [4]P.A.D_16@hotmail.com,
[5]klamath135@gmail.com

N. Iracemi Escalante
Departamento de Ciencias Básicas
Instituto Tecnológico de Pabellón de Arteaga
Carretera a la Estación de Rincón KM 1,
C.P. 20670, Pabellón de Arteaga, Ags., México
Email: aivinsg_2682@hotmail.com

***Abstract*—The aim of this paper is to present two new variations of the frame-level cost function for training a deep neural network in order to achieve better word error rates in speech recognition. Minimization functions of a neural network are salient aspects to deal with when researchers are working on machine learning, and hence their improvement is a process of constant evolution. In the first proposed method, the conventional cross-entropy function can be mapped to a non-uniform loss function based on its corresponding extropy (a complementary dual function), enhancing the frames that have ambiguity in their belonging to specific senones (tied-triphone states in a hidden Markov model). The second proposition is a fusion of the proposed mapped cross-entropy and the boosted cross-entropy function, which emphasizes those frames with low target posterior probability. The developed approaches have been performed by using a personalized mid-vocabulary speaker-independent voice corpus. This dataset is employed for recognition of digit strings and personal name lists in Spanish from the northern central part of Mexico on a connected-words phone dialing task. A relative word error rate improvement of** 12.3% **and** 10.7% **is obtained with the two proposed approaches, respectively, regarding the conventional well-established cross-entropy objective function.**

I. Introduction

Many current speech recognition systems use neural nets in order to find the concordance of input acoustic observations (frames) to an HMM (hidden Markov model) state [1]–[3]. Acoustic modeling with neural networks has shown these can obtain better word error rates (WER) with regard to Gaussian mixture models (GMM-HMM) in most speech recognition tasks under different conditions (e.g., [2], [4]–[10]). Distinct training criteria can be used for optimizing the weights in the DNN-HMM (deep neural network) framework, such as [11]–[13]: i) frame-level training and ii) sequence-discriminative training. It is common to employ cross-entropy (CE) for frame-based training [14], whereas that MMI (maximum mutual information) and MPE/sMBR (minimum phone error / state level minimum Bayes risk) are examples of sequence-based criteria [15]. Few tasks have been focused on exploring frame-level training criteria, and in fact sequence-level training needs a well-trained DNN based on a frame-level criterion [16].

This paper proposes two new variations of the frame-level cost function of a DNN with the purpose of improving error rates in an automatic speech recognition (ASR) system. In the first proposed framework, the *extropy* measure (a complementary dual function of the entropy) [17] is subtracted from CE, transforming it into a variation of itself, termed *non-uniform mapped cross-entropy*. Similar target probabilities from the network output may possess different extropy even though they have the same posterior value. This scheme attempts to eliminate the ambiguity of difficult frames, trying to make more specific their belonging to a senone (a tied context-dependent state). The second proposed method applies to this mapped approach a fusion with the boosted cross-entropy method devised by Huang et al. [16]. In this case, the new formulated loss function emphasizes difficult frames with low target posterior probabilities and deemphasizes the importance of those frames with high DNN prediction. The DNN training frameworks, implemented with the Kaldi toolkit [18], have been applied to a case study in Spanish from the northern central part of Mexico. The ASR task has been performed with a mid-vocabulary corpus on a personalized task of speaker-independent connected words. The best results of the presented models (WER of 2.78% and 2.83%) attain a relative reduction of 12.3% and 10.7% in comparison with the conventional cross-entropy function (3.17% WER).

The rest of the paper is organized as follows. Section II briefly describes the training of a deep neural network using the classical frame-level cost function within a speech

978-1-5386-0819-7/17/$31.00 ©2017 IEEE