




A comparative case study of neural network training by using frame-level cost functions for automatic speech recognition purposes in Spanish

Aldonso Becerra¹  · J. Ismael de la Rosa¹ · Efrén González¹ · A. David Pedroza¹ · N. Iracemi Escalante² · Eduardo Santos¹

Received: 25 January 2019 / Accepted: 18 February 2020 / Published online: 27 March 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Training procedures of a deep neural network are still an area with ample research possibilities and constant improvement either to increase its efficiency or its time performance. One of the lesser-addressed components is its objective function, which is an underlying aspect to consider when there is the necessity to achieve better error rates in the area of automatic speech recognition. The aim of this paper is to present two new variations of the frame-level cost function for training a deep neural network with the purpose of obtaining superior word error rates in speech recognition applied to a case study in Spanish. The first proposed function is a fusion between the boosted cross-entropy and the so called cross-entropy/log-posterior-ratio. The main idea is to jointly emphasize the prediction of difficult/crucial frames provided by a boosting factor and at the same time enlarge the distance between the target senone and its closest competitor. The second proposal is a fusion between the non-uniform mapped cross-entropy and the cross-entropy/log-posterior-ratio. This function utilizes both the mapped function to enhance the frames that have ambiguity in their belonging to specific senones and the log-posterior-ratio with the purpose of separating the target senone against the most competing tied tri-phone state. The proposed approaches are compared against those frame-level cost functions discussed in the state of the art. This comparative has been made by using a personalized mid-vocabulary speaker-independent voice corpus. This dataset is employed for the recognition of digit strings and personal name lists in Spanish from the northern central part of México on a connected-words phone dialing task. A relative word error rate improvement of 15.14% and 12.30% is obtained with the two proposed approaches, respectively, against the plain well-established cross-entropy loss function.

Keywords Speech recognition · Neural networks · Deep learning · Machine learning · Cross-entropy · Frame-level loss function

✉ Aldonso Becerra
a7donso@uaz.edu.mx