Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 316–322

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Improving speech recognition using data augmentation and acoustic model fusion

Ilyes Rebai[a,*], Yessine BenAyed[a], Walid Mahdi[a], Jean-Pierre Lorré[b]

[a] *Multimedia InfoRmation system and Advanced Computing Laboratory*
University of Sfax, Sfax, Tunisia
[b] *Director of Innovation @LINAGORA, Toulouse - France*

Abstract

Deep learning based systems have greatly improved the performance in speech recognition tasks, and various deep architectures and learning methods have been developed in the last few years. Along with that, Data Augmentation (DA), which is a common strategy adopted to increase the quantity of training data, has been shown to be effective for neural network training to make invariant predictions. On the other hand, Ensemble Method (EM) approaches have received considerable attention in the machine learning community to increase the effectiveness of classifiers. Therefore, we propose in this work a new Deep Neural Network (DNN) speech recognition architecture which takes advantage from both DA and EM approaches in order to improve the prediction accuracy of the system. In this paper, we first explore an existing approach based on vocal tract length perturbation and we propose a different DA technique based on feature perturbation to create a modified training data sets. Finally, EM techniques are used to integrate the posterior probabilities produced by different DNN acoustic models trained on different data sets. Experimental results demonstrate an increase in the recognition performance of the proposed system.

Keywords: speech recognition, deep learning, data augmentation, ensemble method, linear logistic regression

1. Introduction

Speech recognition system is mainly based on three models, an acoustic model, a language model and a pronunciation lexicon. Besides, the performance of these models relies greatly on the amount of data used during training. During the last years, several researchers focused their works on improving two key elements in speech recognition: speech data, and acoustic model.

During the past decade, DNN based speech recognition systems have been demonstrated to provide significantly higher accuracy in continuous phone and word recognition tasks than the earlier state-of-the-art GMM-based systems [1,2,3,4]. More recent advances in deep learning techniques, such as convolutional deep neural networks (CNN)

[*] Corresponding author.
E-mail address: rebai_ilyes@hotmail.fr

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.003

and deep recurrent neural networks (RNN), have shown high performance in speech recognition[5,6,7,8]. While deep learning in speech recognition has claimed state-of-the-art performance over conventional GMM based systems, there are sometimes fewer differences in performance between the different deep techniques. Nonetheless, it could be more beneficial to combine these various systems into a single one. In fact, the resulting of the ensemble is generally more accurate than any of the individual model that composes the ensemble[9,10].

Besides, another research direction for speech recognition focused on the development of efficient DA methods. In practice, a large amount of transcribed training data is usually needed to enable accurate speech recognition which is not the case for several languages. Therefore, DA is proposed, where the speech data is artificially augmented by applying different types of transformations. Indeed, it is a common strategy adopted to increase the quantity of training data[11,12,13,14]. For instance, vocal tract length perturbation (VTLP) has shown gains on the TIMIT phoneme recognition task using DNN based acoustic modeling[11].

In this paper, a new DNN based speech recognition system is proposed in which we take advantage from the existing approaches in order to improve the recognition performance. Specifically, DA and EM approaches and are combined into a single system. Indeed, we exploit DA approaches with DNN acoustic modeling. We first explore the VTLP and we propose a different DA technique based on feature perturbation to augment the training data. Next, EM techniques are used to integrate the posterior probabilities produced by different DNN acoustic models trained on different data sets to give improved prediction accuracy. Three types of techniques are evaluated in this work: major voting scheme, average scheme, and Linear Logistic Regression (LLR) for Fusion and Calibration. The experimental results demonstrate the effectiveness of the proposed system.

The remainder of the paper is organized as follows. In Section 2, we give details of the VTLP approach and the proposed feature perturbation technique used to generate transformed input features for deep neural network training. Section 3 reviews the EM and LLR techniques used for acoustic model combination. Next, we present the proposed speech recognition system and the experimental setup used in this work in Section 4 and 5 respectively. Finally, experimental results are presented in Section 6, followed by a conclusion in Section 7.

2. Data Augmentation

Data augmentation is a common strategy adopted to increase the quantity of training data. It is a key ingredient of the state of the art systems for image recognition and speech recognition[15,11]. With the widespread adoption of neural networks in speech recognition systems which require a large speech database for training such a deep architecture, DA is very useful for small data sets. Indeed, it is possible to augment speech databases and to use the augmented database to achieve improved accuracy.

2.1. Vocal Tract Length Perturbation

With DNN based acoustic modeling, vocal tract length perturbation (VTLP)[11] has shown gains on the TIMIT phoneme recognition task. VTLP was further extended to large vocabulary continuous speech recognition (LVCSR)[12]. It was reported that selecting VTLP warping factors from a limited set of perturbation factors was better[12].

In practice, for each utterance in the training set, a warping factor α is randomly chosen from $[0.9, 1.1]$ to warp the frequency axis. Therefore, the vocal tract length of the speaker is slightly perturbed to distort the original speech spectrum of the utterance to create a new replica of it. In this work, a set warping factors, $\{0.9, 1.0, 1.1\}$ is used to create three copies of the original features. Furthermore, the same warping factors are applied to all speakers in the training set.

2.2. Feature Perturbation

We propose in this work a different method for DA and compare it with the existing augmentation technique VTLP. Feature perturbation aims at modifying the extracted acoustic feature vectors by adding random values. This trans-