Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-018-1146-z

3.3.3 (3)

ORIGINAL RESEARCH

Multi-level region-of-interest CNNs for end to end speech recognition

Shubhanshi Singhal[1] · Vishal Passricha[2] · Pooja Sharma[3] · Rajesh Kumar Aggarwal[2]

Received: 8 May 2018 / Accepted: 23 November 2018

Abstract
Efficient and robust automatic speech recognition (ASR) systems are in high demand in the present scenario. Mostly ASR systems are generally fed with cepstral features like mel-frequency cepstral coefficients and perceptual linear prediction. However, some attempts are also made in speech recognition to shift on simple features like critical band energies or spectrogram using deep learning models. These approaches always claim that they have the ability to train directly with the raw signal. Such systems highly depend on the excellent discriminative power of ConvNet layers to separate two phonemes having nearly similar accents but they do not offer high recognition rate. The main reason for limited recognition rate is stride based pooling methods that performs sharp reduction in output dimensionality i.e. at least 75%. To improve the performance, region-based convolutional neural networks (R-CNNs) and Fast R-CNN were proposed but their performances did not meet the expected level. Therefore, a new pooling technique, multilevel region of interest (RoI) pooling is proposed which pools the multilevel information from multiple ConvNet layers. The newly proposed architecture is named as multilevel RoI convolutional neural network (MR-CNN). It is designed by simply placing RoI pooling layers after up to four coarsest layers. It improves extracted features using additional information from the multilevel ConvNet layers. Its performance is evaluated on TIMIT and Wall Street Journal (WSJ) datasets for phoneme recognition. Phoneme error-rate offered by this model on raw speech is 16.4% and 17.1% on TIMIT and WSJ datasets respectively which is slightly better than spectral features.

Keywords Cepstral features · End-to-end training · Feature extraction · Pooling · Raw speech · Spectral features

1 Introduction

In traditional ASR systems, feature extraction and acoustic modeling are done in two separate steps. Mostly speech recognition systems divide the task into several subtasks. In the first step, features are extracted from the speech signals based on the task-specific knowledge of the phenomena. Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein 1990) and perceptual linear prediction (PLP) (Hermansky 1990) are the popular cepstral based feature extraction techniques. In the second step, the conditional probabilities of phonemes are calculated using either generative or discriminative models. Finally, sequence under constraints is recognized using dynamic programming techniques. Although some advanced convolutional neural network (CNN) based systems perform both feature learning and acoustic modeling in a single step as shown in Fig. 1. In these systems, feature learning and acoustic modeling are performed by CNN and decoding task is completed using hidden Markov model (HMM). Abdel-Hamid et al. (2012) applied CNNs concepts in frequency domain using mel-frequency spectral coefficients (MFSC) (i.e. with no discrete cosine transformation) to learn efficient feature representation by normalizing acoustic variations and achieved improvement in phoneme recognition rate.

Efficient feature extraction has been highly demanded in speech recognition research. The existing feature extraction methods are divided into two broad categories i.e. cepstral based methods and raw waveform based methods. Methods in the first category like MFCC and PLP are based on sliding windows mechanism. They are accurate enough but have high computational cost and time. However, model in the second category like deep learning models directly learn the features. In this category, systems are directly fed with raw speech signals

✉ Rajesh Kumar Aggarwal
rka15969@gmail.com

1 Technology Education and Research Integrated Institute, Kurukshetra, India

2 National Institute of Technology, Kurukshetra, India

3 Government College for Women, Karnal, India

Published online: 29 November 2018