Neurocomputing 106 (2013) 148–157

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Exploiting deep neural networks for detection-based speech recognition

Sabato Marco Siniscalchi [a,b,*], Dong Yu [c], Li Deng [c], Chin-Hui Lee [b]

[a] *Faculty of Engineering and Architecture, Kore University of Enna, Cittadella Universitaria, Enna, Sicily, Italy*
[b] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*
[c] *Speech Research Group, Microsoft Research, Redmond, WA, USA*

ARTICLE INFO

Article history:
Received 29 May 2012
Received in revised form
21 September 2012
Accepted 5 November 2012
Communicated by R. Capobianco Guido
Available online 23 November 2012

Keywords:
Deep neural networks
Multi-layer perceptrons
Articulatory recognition
Speech recognition
Lattice rescoring

ABSTRACT

In recent years deep neural networks (DNNs) – multilayer perceptrons (MLPs) with many hidden layers – have been successfully applied to several speech tasks, i.e., phoneme recognition, out of vocabulary word detection, confidence measure, etc. In this paper, we show that DNNs can be used to boost the classification accuracy of basic speech units, such as phonetic attributes (phonological features) and phonemes. This boosting leads to higher flexibility and has the potential to integrate both top-down and bottom-up knowledge into the Automatic Speech Attribute Transcription (ASAT) framework. ASAT is a new family of lattice-based speech recognition systems grounded on accurate detection of speech attributes. In this paper we compare DNNs and shallow MLPs within the ASAT framework to classify phonetic attributes and phonemes. Several DNN architectures ranging from five to seven hidden layers and up to 2048 hidden units per hidden layer will be presented and evaluated. Experimental evidence on the speaker-independent Wall Street Journal corpus clearly demonstrates that DNNs can achieve significant improvements over the shallow MLPs with a single hidden layer, producing greater than 90% frame-level attribute estimation accuracies for all 21 phonetic features tested. Similar improvement is also observed on the phoneme classification task with excellent frame-level accuracy of 86.6% by using DNNs. This improved phoneme prediction accuracy, when integrated into a standard large vocabulary continuous speech recognition (LVCSR) system through a word lattice rescoring framework, results in improved word recognition accuracy, which is better than previously reported word lattice rescoring results.

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems often rely on a pattern matching framework that expresses spoken utterances as sequences of stochastic patterns [1]. Top-down approaches are usually adopted to represent all constraints in a single, compact probabilistic finite state network (FSN), composed of acoustic hidden Markov model (HMM) states with emission probabilities generated by Gaussian mixture models (GMMs), phones, lexicon, grammar nodes, and their connecting arcs [2]. For a given input utterance, the maximum a posteriori decoding [1] procedure is used to find the most possible sequence of words embedded in the FSN as the recognized sentence. This search technique, known as the top-down integrated search strategy, has attained remarkable results in many ASR tasks. Nonetheless, recognition error rates for difficult tasks, such as spontaneous and unconstrained speech recognition, are still unacceptably high. In contrast, there is evidence to show that bottom-up, stage-by-stage ASR paradigm may do better under some spontaneous speech phenomena [3]. Automatic speech attribute transcription (ASAT) [4], a promising alternative ASR paradigm, is a bottom-up framework that first detects a collection of speech attribute cues and then integrates such cues to make linguistic validations. A typical ASAT system uses the articulatory-based phonological features studied earlier [5–9] in a new detection-based framework. ASAT has been extended and applied to a number of tasks including rescoring of word lattices generated by state-of-the-art HMM systems [10], continuous phoneme recognition [11], cross-language attribute detection and phoneme recognition [12] and spoken language recognition [13]. The speech cues detected in ASAT are referred to as *speech attributes*. The terms phonological features and speech attributes will be used interchangeably in this work.

In recent years there has also been a considerable resurgence of interest in neural network approaches to speech recognition. Neural networks are powerful pattern recognition tools that have been used for several real world applications [14], and different

[*] Corresponding author at: Faculty of Engineering and Architecture, Kore University of Enna, Cittadella Universitaria, Enna, Sicily, Italy. Tel.: +39 3472913375.
E-mail addresses: marco.siniscalchi@unikore.it (S.M. Siniscalchi), dongyu@microsoft.com (D. Yu), deng@mircosoft.com (L. Deng), chl@ece.gatech.edu (C.-H. Lee).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.008

successful techniques have been developed around them since the early ‘80s in the speech community. For example, in connectionist speech recognition systems [15], neural networks are used to estimate the state emission probabilities of a HMM. In the TANDEM approach [16], a neural network extracts discriminative speech features that are fed into conventional GMM-HMM-based speech recognizers. In detection-based ASR paradigms (e.g., [11]), a set of neural networks learns the mapping from a spectral-based feature space to a phonetic feature space. Neural networks have also been used to model state and transition features in conditional random field (CRF) based ASR systems (e.g., [17]), in beam search pruning [18] and confidence measure estimation [19,20]. Although several architectures have been proposed to tackle different speech recognition tasks, such as recurrent neural networks (e.g., [21,22]) and time-delay neural network [23], the stylistic characteristics of the MLPs is by far the most popular due to the compromise realized between recognition rate, recognition speed, and memory resources. Furthermore, it has been shown that feed-forward neural architectures can approximate any function defined on compact sets in $\mathbf{R}^n$ [24], that is, they are *universal approximators* [14].

More recently, a major advance has been made in training densely connected, generative deep belief nets (DBNs) with many hidden layers. The core idea of the DBN training algorithm suggested in [25] is to first initialize the weights of each layer greedily in a purely unsupervised way by treating each pair of the layers as a restricted Boltzmann machine (RBM) and then fine-tune all the weights jointly to further improve the likelihood. The resulting DBN can be considered as a hierarchy of nonlinear feature detectors that can capture complex statistical patterns in data. For classification tasks, the same DBN pre-training algorithm can be used to initialize the weights in deep neural networks (DNNs) – MLPs with many hidden layers. The weights in the entire DNN can then be fine-tuned using labeled data. DNNs have been proven to be effective in a number of applications, including coding and classification of speech, audio, text, and image data [26–30]. These advances triggered interest in developing acoustic models based on DNNs and other deep learning techniques for ASR. For example, the context-independent DNN-HMM hybrid architectures have recently been proposed for phoneme recognition [31,32] and have achieved very competitive performance. A novel acoustic model, the context-dependent (CD)-DNN-HMM proposed in [33] has been successfully applied to large vocabulary speech recognition tasks and can cut word error rate by up to one third on the challenging conversational speech transcription tasks compared to the discriminatively trained conventional CD-GMM-HMM systems [34].

In this study[1], elements of both of these two research directions, namely ASAT and DNN, are merged together, and the conventional shallow MLPs used in [36] are replaced with DNNs, which has been shown to have very good theoretical properties [37] and demonstrated superior performances for both phoneme [31,32] and word recognition [33,34,38,39]. Following the ASAT paradigm, a *bank of speech attribute detectors* that assign probabilistic scores to manner and place of articulation events is built using DNNs. Then a DNN is designed to (1) combine together the output of these detectors and (2) generate phoneme posterior probability. A wide range of DNN architectures will be built by extending the conventional single hidden layer MLPs to five and seven layers. Experimental evidence on the speaker independent Wall Street Journal dataset [40] demonstrates that the proposed solution outperforms conventional shallow MLPs in both attribute and phoneme classification. Furthermore, by re-scoring the set of most likely hypotheses embedded in the word lattices generated by a conventional HMM-based LVCSR system using the DNN phoneme posterior probabilities, a two-stage LVCSR recognizer gave relative word error rate (WER) reductions ranging from 8.7% to 13.0% over the initial result and improves over previous studies on word lattice rescoring [10].

This result along with the significantly boosted quality in attribute and phoneme estimation makes it highly promising to advance *bottom-up* LVCSR with DNNs and with new ways of incorporating the key asynchrony properties of the articulatory-motivated phonetic attributes. This also opens doors to new flexibility in combining top-down and bottom-up ASR. Furthermore, it should be noted that modeling of articulatory-based phonetic features and phoneme is an active research filed in automatic speech recognition. Therefore, the current investigation can also impact research areas beyond the ASAT framework. For instance, several researchers have argued that better results can be achieved by modeling the underlying processes of co-articulation and assimilation rather than simply describing their effects on the speech signal (e.g., [7]). It is also believed that by integrating articulatory-motivated information into the speech recognition engine most of the problems of the current technology can be addressed. Finally, phoneme estimation also plays a very important role in many speech processing applications, such as out-of-vocabulary detection [41] and language identification (e.g., [42]).

The remainder of the paper is organized as follows. A brief survey on the ASAT paradigm for speech recognition is given in Section 2. Section 3 gives a light overview of related works on articulatory-motivated phonological attributes and phoneme estimation. The DNN architecture and training scheme are discussed in Section 4. The word lattice rescoring procedure adopted in this study is outlined in Section 5. Next, the experimental setup is given in Section 6 in which experimental results on attributes and phoneme classification, and word lattice rescoring are presented and discussed. Finally, we discuss our findings and conclude our work in Section 7.

2. ASAT in a nutshell

It is well known that the speech signal contains a rich set of information that facilitates human auditory perception and communication, beyond a simple linguistic interpretation of the spoken input. In order to bridge the performance gap between ASR systems and human speech recognition (HSR), the narrow notion of speech-to-text in ASR has to be expanded to incorporate all related information “embedded” in speech utterances. This collection of information includes a set of fundamental speech sounds with their linguistic interpretations, a speaker profile encompassing gender, accent, emotional state and other speaker characteristics, the speaking environment, etc. Collectively, we call this superset of speech information the attributes of speech. They are not only critical for high performance speech recognition, but also useful for many other applications, such as speaker recognition, language identification, speech perception, speech synthesis, etc. ASAT therefore promises to be knowledge-rich and capable of incorporating multiple levels of information in the knowledge hierarchy into attribute detection, evidence verification and integration, i.e., all modules in the ASAT system [4]. Since speech processing in ASAT is highly parallel, a collaborative community effort can be built around a common sharable platform to enable a ‘‘divide-and-conquer’’ ASR paradigm that facilitates tight coupling of interdisciplinary studies of speech science and speech processing [4]. A block diagram of the ASAT approach to ASR is shown in Fig. 1. The top panel shows the general ASAT front end that performs a collection of speech analyses geared to

[1] This work re-organizes, expands, and completes our study reported in [35].